

BioChem: the Fisheries and Oceans Canada Database for Biological and Chemical Data

L. Devine¹, M.K. Kennedy², I. St-Pierre¹, C. Lafleur¹,
M. Ouellet³, and S. Bond²

Science Branch
Fisheries and Oceans Canada

¹ Institut Maurice-Lamontagne
850, route de la Mer
Mont-Joli, QC G5H 3Z4

² Bedford Institute of Oceanography
1 Challenger Drive
Dartmouth, NS B2Y 4A2

³ Oceanography and Scientific Data
200 Kent Street
Ottawa, ON K1A 0E6

2014

**Canadian Technical Report of
Fisheries and Aquatic Sciences 3073**



Pêches
et Océans

Fisheries
and Oceans

Canada 

Canadian Technical Report of Fisheries and Aquatic Sciences

Technical reports contain scientific and technical information that contribute to existing knowledge but that are not normally appropriate for primary literature. Technical reports are directed primarily toward a worldwide audience and have an international distribution. No restriction is placed on subject matter, and the series reflects the broad interests and policies of the Department of Fisheries and Oceans, namely, fisheries and aquatic sciences.

Technical reports may be cited as full publications. The correct citation appears above the abstract of each report. Each report is indexed in the data base *Aquatic Sciences and Fisheries Abstracts*.

Numbers 1-456 in this series were issued as Technical Reports of the Fisheries Research Board of Canada. Numbers 457-714 were issued as Department of the Environment, Fisheries and Marine Service, Research and Development Directorate Technical Reports. Numbers 715-924 were issued as Department of Fisheries and the Environment, Fisheries and Marine Service Technical Reports. The current series name was changed with report number 925.

Technical reports are produced regionally but are numbered nationally. Requests for individual reports will be filled by the issuing establishment listed on the front cover and title page. Out-of-stock reports will be supplied for a fee by commercial agents.

Rapport technique canadien des sciences halieutiques et aquatiques

Les rapports techniques contiennent des renseignements scientifiques et techniques qui constituent une contribution aux connaissances actuelles, mais qui ne sont pas normalement appropriés pour la publication dans un journal scientifique. Les rapports techniques sont destinés essentiellement à un public international et ils sont distribués à cet échelon. Il n'y a aucune restriction quant au sujet; de fait, la série reflète la vaste gamme des intérêts et des politiques du ministère des Pêches et des Océans, c'est-à-dire les sciences halieutiques et aquatiques.

Les rapports techniques peuvent être cités comme des publications intégrales. Le titre exact paraît au-dessus du résumé de chaque rapport. Les rapports techniques sont indexés dans la base de données *Aquatic Sciences and Fisheries Abstracts*.

Les numéros 1 à 456 de cette série ont été publiés à titre de rapports techniques de l'Office des recherches sur les pêcheries du Canada. Les numéros 457 à 714 sont parus à titre de rapports techniques de la Direction générale de la recherche et du développement, Service des pêches et de la mer, ministère de l'Environnement. Les numéros 715 à 924 ont été publiés à titre de rapports techniques du Service des pêches et de la mer, ministère des Pêches et de l'Environnement. Le nom actuel de la série a été établi lors de la parution du numéro 925.

Les rapports techniques sont produits à l'échelon régional, mais numérotés à l'échelon national. Les demandes de rapports seront satisfaites par l'établissement d'origine dont le nom figure sur la couverture et la page du titre. Les rapports épuisés seront fournis contre rétribution par des agents commerciaux.

Canadian Technical Report of
Fisheries and Aquatic Sciences 3073

2014

BioChem:
the Fisheries and Oceans Canada Database
for Biological and Chemical Data

L. Devine¹, M.K. Kennedy², I. St-Pierre¹, C. Lafleur¹,
M. Ouellet³, and S. Bond²

Science Branch
Fisheries and Oceans Canada

¹Institut Maurice-Lamontagne
850, route de la Mer
Mont-Joli, QC G5H 3Z4

²Bedford Institute of Oceanography
1 Challenger Drive
Dartmouth, NS B2Y 4A2

³Oceanography and Scientific Data
200 Kent Street
Ottawa, ON K1A 0E6

© Her Majesty the Queen in Right of Canada, 2014
Cat. No. Fs 97-6/3073E-PDF ISSN 1488-5379

Correct citation for this publication:

Devine, L., M.K. Kennedy, I. St-Pierre, C. Lafleur, M. Ouellet, and S. Bond. 2014. BioChem:
the Fisheries and Oceans Canada database for biological and chemical data. Can. Tech. Rep.
Fish. Aquat. Sci. 3073: iv + 40 pp.

TABLE OF CONTENTS

Abstract	iv
Résumé.....	iv
Preface.....	v
1.0 Introduction.....	1
1.1 History and overview	1
1.2 BioChem data sources.....	1
1.3 BioChem missions	3
1.4 DFO data inventory and dataset collections	4
2.0 The BioChem database	4
2.1 BioChem database design	4
2.2 BioChem functional areas: discrete and plankton	6
2.3 BioChem code tables	9
2.4 Quality control	11
3.0 Using BioChem.....	14
3.1 Retrieving data using the query application.....	15
3.2 Loading data using the edit application	16
4.0 BioChem and global data initiatives	16
5.0 Acknowledgements.....	18
6.0 References	18
Annex I. List of acronyms and abbreviations used in this document.	20
Annex II. Mission descriptors.....	22
Annex III. Plankton collection methods	24
Annex IV. BioChem taxonomic code table	26
Annex V. How to interpret BioChem query output.....	28
Annex VI. BioChem load tables	32

ABSTRACT

Devine, L., M.K. Kennedy, I. St-Pierre, C. Lafleur, M. Ouellet, and S. Bond. 2014. BioChem: the Fisheries and Oceans Canada database for biological and chemical data. Can. Tech. Rep. Fish. Aquat. Sci. 3073: v + 40 pp.

BioChem is the name given to the database developed and maintained by Fisheries and Oceans Canada to hold biological and chemical data resulting from department research initiatives or concerning areas of Canadian interest. The data held in BioChem are divided into two functional areas: discrete data (usually from water bottle sampling; examples are nutrients, dissolved oxygen, and chlorophyll) and plankton data (usually from towed nets; examples are species counts and biomass measures). This report gives an overview of BioChem's history, design, and data holdings (as of the report's publication); provides links to key code tables; describes the quality control procedures applied to data before archive; supplies explanations of how to build and interpret queries; and outlines the procedure necessary to load data to the archive. A discussion of how BioChem fits with global initiatives is also included.

RÉSUMÉ

Devine, L., M.K. Kennedy, I. St-Pierre, C. Lafleur, M. Ouellet, and S. Bond. 2014. BioChem: the Fisheries and Oceans Canada database for biological and chemical data. Can. Tech. Rep. Fish. Aquat. Sci. 3073: v + 40 pp.

BioChem est le nom de la base de données développée et gérée par Pêches et Océans Canada pour archiver les données biologiques et chimiques résultant des initiatives de recherche du ministère, ou concernant des domaines d'intérêt canadien. Les données contenues dans BioChem sont divisées en deux domaines fonctionnels : les données discrètes (généralement échantillonnées à l'aide d'une bouteille; par exemple, les sels nutritifs, l'oxygène dissous et la chlorophylle) et les données de plancton (généralement échantillonnées à partir de filets, comme par exemple le dénombrement d'espèces et la mesure de la biomasse). Ce rapport donne un aperçu de l'historique de BioChem et de sa conception, ainsi qu'une idée de la quantité de données qui s'y trouvent (au moment de la publication du rapport). De plus, il contient des liens vers les tables de codes; décrit les procédures de contrôle de qualité appliquées aux données avant l'archivage; fournit des explications sur la façon de créer et d'interpréter les requêtes; et détaille la procédure nécessaire pour charger les données dans l'archive. La façon dont BioChem cadre avec les initiatives mondiales est également discutée.

PREFACE

It is challenging to write a document describing the conception, evolution, and current state of an active database. In this report, we give the background circumstances that led to BioChem's creation, make note of the major phases it went through, and describe its present status. We describe data sources and give a summary of the amount and types of data currently held therein. An overview of the database's design is included along with sections that give information on how to query the database and how to interpret the query results; an additional section includes the methods used by data managers to archive data. Finally, we present a brief section on how the BioChem database fits into global data initiatives. Since BioChem is an active database, it will undoubtedly undergo further changes that may make sections of this report obsolete in the future. Users will be made aware of critical changes by updates to the BioChem website (<http://www.meds-sdmm.dfo-mpo.gc.ca/biochem/biochem-eng.htm>).

1.0 INTRODUCTION

1.1 HISTORY AND OVERVIEW

The creation of a national Fisheries and Oceans Canada (DFO) archive of marine biological and chemical data has been the long-term goal of many dedicated individuals (see Gregory and Narayanan 2003 for an early overview of the project). The initial idea for the project was raised during a 1997 meeting of DFO's Climate Science Coordinating Committee, when scientists reported that one fundamental problem was the lack of biological databases to assess the implications of climate change in developing monitoring programs. That same year, a national group met in Ottawa to develop the framework for a national biological / chemical database. This was followed in 1998 with workshops at the Bedford Institute of Oceanography (BIO) and Marine Environmental Data Service (MEDS; known as Oceanography and Scientific Data branch [OSD] since 2013 and had been Integrated Science Data Management from 2006 to 2013) to finalize the initial design. The design was presented at the November 1998 meeting of the Atlantic Zone Monitoring Program (AZMP) and was adopted as a model for data generated by this monitoring program.

The project's next major advance occurred when BIO's Ocean Sciences and Marine Environmental Sciences divisions decided that the database's development in Maritimes region would be included as a priority within their Y2K preparedness. Other sources of funding included Maritimes region's AZMP.

Over the next four years, the development of the BioChem database, as it came to be called, was led by Maritimes region through a partnership of BIO data managers and database developers in the Maritimes region informatics group. While development of the application was one challenge, the migration of decades of biological and chemical data observations into the database has been an even greater one.

In 2003, BioChem was migrated to MEDS (now OSD) in Ottawa to finally achieve its initial goal as a national resource accessible from all regions. From 2006 onward, a steering committee was established under DFO's National Science Data Management Committee to—among other things—oversee the operation of BioChem. BioChem was initially accessible only to DFO personnel via the DFO intranet; it was made available on the internet starting in 2008, thus open for use by the scientific community and general public.

A list of the acronyms and abbreviations used in this document can be found in Annex I.

1.2 BIOCHEM DATA SOURCES

The BioChem archive mainly includes data from current DFO programs and legacy datasets; these are loaded to the archive by regional data managers. When the Maritime's regional version

of the database first came online in June 2000, it was populated with records (discrete data¹) from BIO's Marine Chemistry Division databases, which included nutrient, temperature, and salinity data collected by DFO researchers and augmented with data extracted from NOAA's World Ocean Database (WOD94 and WOD98) for the Northwest Atlantic (35–80°N; 42–100°W) and also with plankton and discrete data belonging to researchers from the Biological Oceanography Division at BIO. A subset of OSD's historical bottle archive has also been loaded. This bottle archive was created in the 1960s as a national repository of observations of water properties collected using oceanographic water bottles from the primary area of Canadian interest (35–90° N; 40–180° W) as well as any other data collected by Canadian scientists anywhere in the world. The remainder of the bottle archive will be migrated to BioChem as time allows.

BioChem is the data archive for bottle and plankton data resulting from AZMP; since 1999, this program has been the source of a large proportion of data recently loaded to the archive. Other data sources include international collaborative programs, universities, and the Sir Alister Hardy Foundation for Ocean Science (SAHFOS, Plymouth, UK; continuous plankton recorder [CPR]).

The database has continued to grow, and as of December 2013 contained records from more than 2100 missions from locations around the globe (Fig. 1). Figure 2 shows the number of archived missions by sampling year: the earliest discrete data record as of December 2013 dates from 1921 and plankton from 1914.

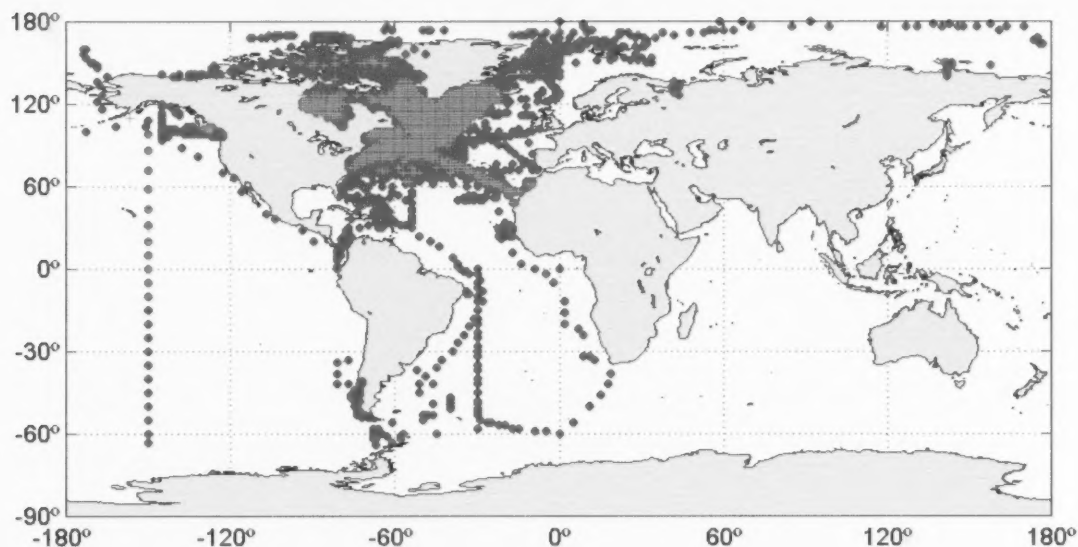


Figure 1. BioChem data holdings: map showing sampling locations (as of December 2013). Blue circles: discrete data; red crosses: plankton data. Sampling positions are binned to one degree squares.

¹ Discrete data result from samples taken at a fixed location in time and space. In BioChem, they generally result from a water bottle sample (so are often called "bottle data"), but they can also be from, e.g., a sediment grab or ice core.

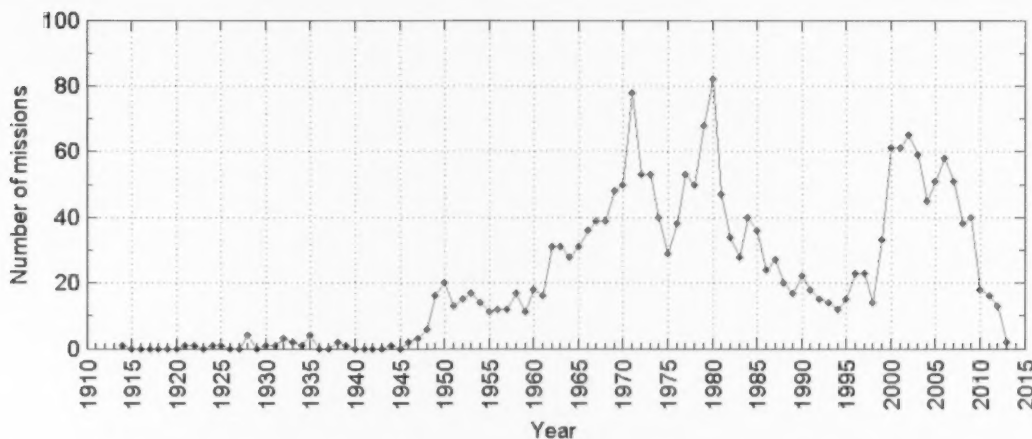


Figure 2. BioChem data holdings: number of missions per year (total of > 2100 as of December 2013). This figure is interesting in that it shows peaks resulting from increases in government spending on research programs as well as large international projects. The increase in the late 1990s is largely due to the start of the Atlantic Zone Monitoring Program (AZMP). The sharp drop at the end is an artifact resulting from the lag between sample collection and data archive.

1.3 BIOCHEM MISSIONS

In the BioChem database, the term “mission” is used to refer to a group of sampling events; these may have been made from one platform or several platforms participating in the same study. For example, a mission could be a multidisciplinary oceanographic cruise from a large ship, nearshore sampling from a smaller vessel, or a collection of sampling events such as weekly visits to a station over the sampling season. It may also happen that all of the data received from an institute for a given year—and even sometimes across more than one year—are considered as a single mission. Thus even though there are a variety of definitions for a mission, the primary function is to identify the data collection so that it is easy to recognize and tracking its history is simplified.

In recent years, all datasets loaded to BioChem have been issued a unique mission descriptor that allows their identification. These mission descriptors are assigned by OSD to assure their unique quality, to allow coordination among regions, and to standardize their format (Annex II). It is unfortunate that the use of these unique identifiers was not initiated when data loading began. Because of this, some datasets have inadvertently been loaded more than once. While personnel are working to eliminate these replications, users should be aware of their existence when considering query results.

As part of its mandate, OSD acquires data of Canadian interest from other countries, and some of these data are included in BioChem. The mission descriptors are kept as-is, and most use the same convention for the first four characters (the platform code, which begins with a two-character country code). However, in many cases the rest of the code does not follow the usual OSD format.

1.4 DFO DATA INVENTORY AND DATASET COLLECTIONS

As part of its national science data management strategy, DFO has created an online inventory of its datasets called DFO MEST ("metadata entry and search tool"). The objective is to make DFO data collections discoverable and accessible. Inventory entries have been created for BioChem as a whole and also for series or collections of related datasets. These related datasets have some commonality that allows them to be grouped—usually they belong to a single program that used common sampling and analysis protocols.

Metadata pages describe collections and include information on the geographic and temporal scope of the data. They also include keywords to facilitate discovery, citation and contact information, a dataset summary and purpose, and a list of publications that used or may help with the interpretation of the data. These pages should provide information related to sample collection and analysis procedures that are not archived along with the data in the database. Having this information available will give BioChem users access to further insight that will help them to correctly interpret their query results. A good summary should provide enough information to assist with decisions related to fitness of use for different research purposes.

DFO MEST is currently only accessible via the DFO intranet, but there are plans to make bilingual collection level records accessible. At that time, a link will be added to the BioChem internet page and query application.

2.0 THE BIOCHEM DATABASE

2.1 BIOCHEM DATABASE DESIGN

In the past, data reports were used to record details of research programs; these reports included the program rationale, sampling methodology, analysis methods, and often the raw data. Starting in the mid-1980s, the number of such reports created started to diminish as data were made digitally accessible. The BioChem database was designed to store information that was formerly provided in these data reports. Decisions were made as to what information had to be included to properly describe the methodology and to interpret the data. Data records in BioChem are associated with a set of relational tables containing information about sample collection such as date, location information (latitude, longitude, sounding, locality, or place name), sampling gear, sampling depth, and so on. BioChem's entity relationship diagrams (ERD) are available on the website (<http://www.meds-sdmm.dfo-mpo.gc.ca/biochem/private-prive/documents-eng.htm>); Figure 3 shows a simplified version of the database structure. When BioChem was first conceived, the intention was that it would accommodate all types of marine data. At that time, these were divided into the broad categories of plankton, discrete (water samples), continuous (profilers, flow-through systems), meteorological, and large binary objects (e.g., satellite images or videos). These different categories were termed "functional areas" and were defined based on the type of ancillary information required to document them. At present, BioChem has two operational functional areas: discrete and plankton.

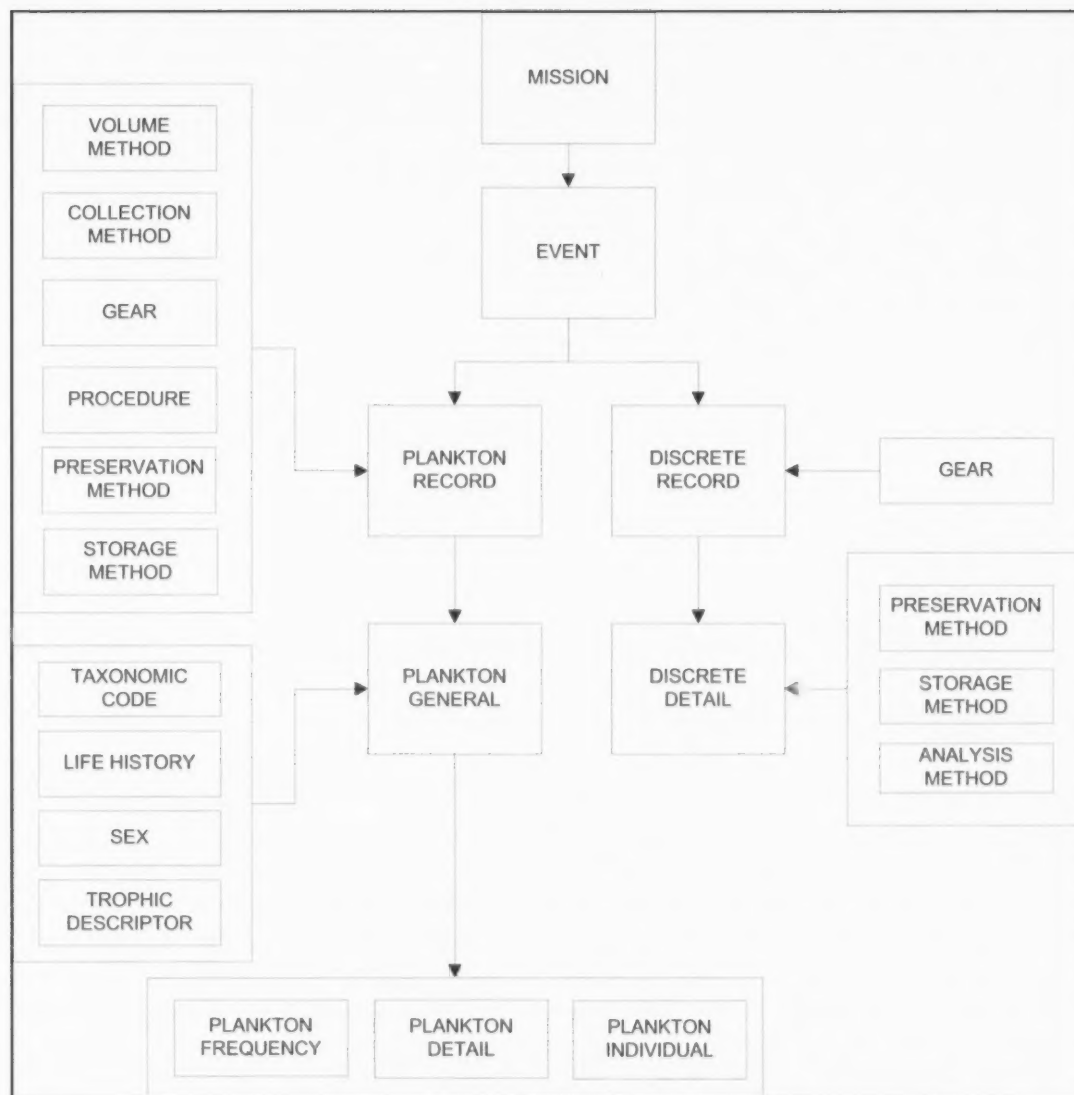


Figure 3. Simplified schematic representation of BioChem's database structure.

The BioChem database comprises three hierarchical levels of metadata followed by the data. Metadata are "data about data," i.e., information that answer the when, where, who, and how types of questions about the data. These levels reflect the generally oceanographic origin of the data. The data themselves are presented as data type / data value; also included are any other information vital to their correct interpretation.

1. Mission level. Mission metadata include information common to the whole mission, such as the cruise identification number (the OSD mission descriptor), chief scientist's name and institute, mission start and end date, sampling platform (usually ship name), and an indication of the general geographic area where sampling took place.

2. Event level. A mission is made up of a collection of events, the definition of which is left to the chief scientist or data custodian; not all groups from the different DFO regions define event in the same way. An event is most easily understood as data that "go together"; specific time and space limits are commonly used to provide linkage between different sampling activities that share a common bond. For example, if bottles, CTD, and nets were deployed at the same station and are intended to be considered together, they would have the same event-level information. Data from the same event may be considered together for interpretation, analysis, or calibration. The most important metadata at the event level include start/end times and minimum and maximum positions along with the event ID.
3. Record level. At this level, data from the two functional areas (discrete and plankton) start to diverge in metadata requirements. While both include start/end times and positions as well as sample ID and sampling gear used, plankton record-level metadata are more detailed. They include information on how the samples were collected, preserved, stored, and analyzed. In addition, information on how the gear was deployed (e.g., vertical or horizontal tows; see Annex III for collection methods), how the volume of water filtered by a net was calculated, and sample processing details such as the level or status of sample analysis.
4. Data level. Data level records contain the data type and the data value.
 - For discrete data, the different data types are defined based on the sample handling, preservation, storage, and analysis methods. All data from a given variable are archived in the same units. The database retains information on the original measured units and the equations used for unit conversion.
 - For plankton data, the data type definition is a combination of the following codes: taxonomic name, life history stage, sex, trophic level, min/max sieve, and name modifier. Plankton data values may be quantitative or relative abundances, biomass (wet or dry weights), settled volume measurements, or simply a qualitative observation (i.e., presence/absence). The database was designed to store raw data values, and information is provided to allow the conversion to standardized units such as number or weight per cubic metre or square metre (integrated water-column values).

The database includes many code tables that detail sampling gear, collection methods, preservation, storage, analyses, and so on (Table 1).

2.2 BIOCHEM FUNCTIONAL AREAS: DISCRETE AND PLANKTON

As already mentioned, the data held in BioChem fall into one of the two functional areas: those collected from discrete samples, i.e., at a specific location, time, and depth (typically water bottle samples) and those associated with biological classification (typically plankton samples). Most discrete data records in BioChem are related to water chemistry and grouped into classes such as nutrients, pigments, organic matter, and oxygen (Fig. 4). Both phytoplankton and zooplankton data occur in the database, although most records are zooplankton (primarily arthropods [crustaceans] and chordates [fish larvae]) (Fig. 5). Physical variables such as temperature and

salinity are also included as ancillary data when available, but BioChem is not the primary repository for these types of data.

2.2.1 Discrete functional area

Discrete data are entered in the database as data type / data value. The data type definition is based on the methods used to process, preserve, store, and/or analyze the sample. Each data type is grouped with similar data types into a category known as the "data retrieval type." For example, chlorophyll *a* is one data retrieval type that currently comprises nine different data types (which are largely based on differences in the analysis method used). Grouping data types into data retrieval types enforces standardized archival and output formats (units, precision [number of decimal points], min/max value range). Data retrievals are further grouped into "classes" to facilitate queries. For example, the "pigments" class includes chlorophylls a, b, c; phaeopigments; and less abundant pigments determined by HPLC analysis; the "nutrients" class includes nitrites, nitrates, phosphates, silicates, ammonia, and urea.

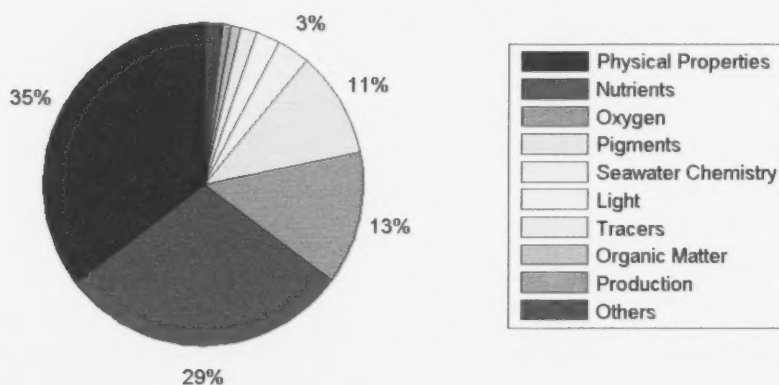


Figure 4. BioChem data holdings: chart showing percentages of the most common discrete data types grouped by class (as of December 2013). While BioChem is not the main repository for physical data (e.g., temperature and salinity), they make up the largest proportion of discrete data in the database because these data are nearly always collected any time any other discrete data are collected.

2.2.2 Plankton functional area

Plankton data records include information on the data type (taxonomic name, life history, sex, trophic level, min/max sieve, and name modifier) along with data values that could include abundance, percentage, biomass, settled volume, and/or an indication of presence-absence.

There are more than 5400 entries in the BioChem taxonomic code table (as of December 2013). Like most species lists, this table includes many names that may not refer to an individual species (e.g., non-taxonomic groups like invertebrates, shrimp, fish, or worms; groupings of species or life history stages that are difficult to distinguish) and may not even refer to living biota (e.g., oil, plastic, sand).

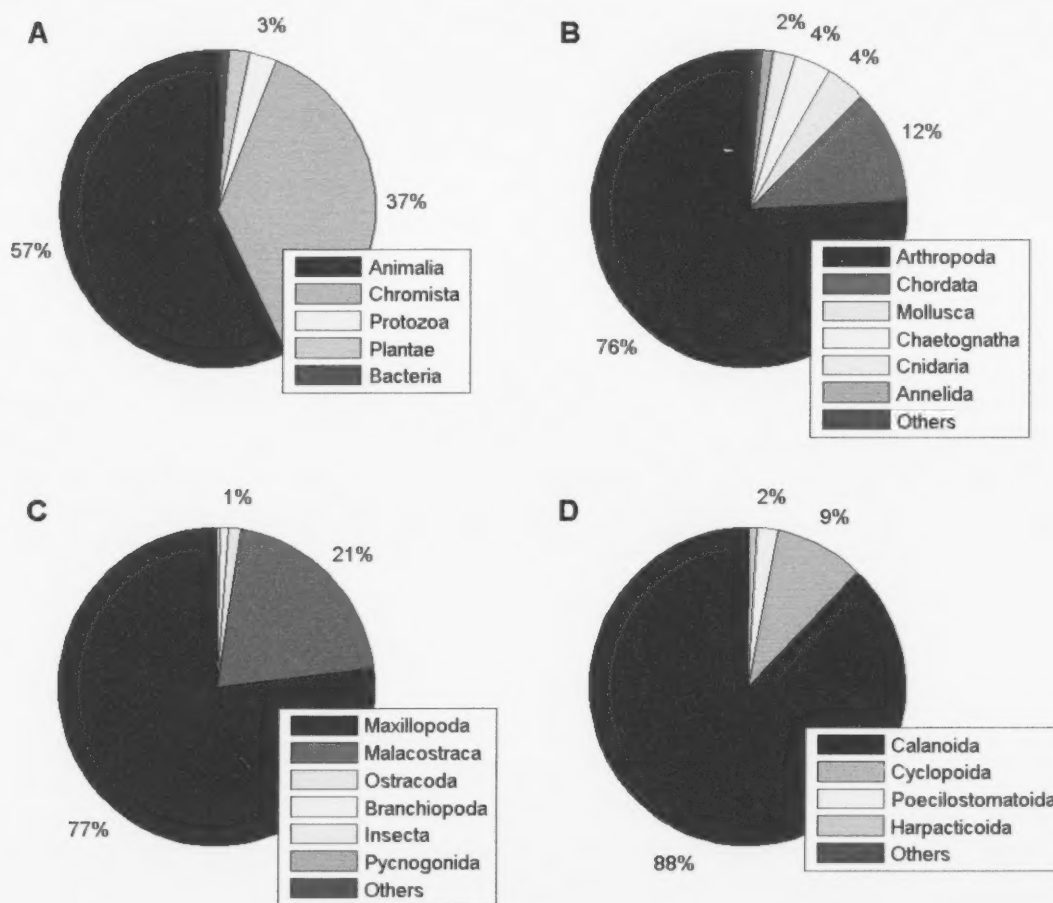


Figure 5. BioChem data holdings: charts showing percentages of the plankton data (as of December 2013). A: kingdoms; B: phyla within Animalia; C: classes within Arthropoda; D: orders within Maxillopoda.

The recommended data management practice to resolve these known issues is to map all records to valid names referencing a recognized standard (Kennedy and Bajona 2009). When the database was first created, names were associated with NODC (US National Oceanographic Data Center, NOAA) codes, and BioChem data managers followed that group's practice of assigning negative codes to names that were not on the NODC list. The NODC code list was abandoned in

the mid-1990s in favour of the new Integrated Taxonomic Information System (ITIS), which uses TSNs ("taxonomic serial number") (see Annex IV for a discussion of taxonomic codes used in BioChem).

However, the main focus of ITIS is not marine, so many valid taxonomic groups in BioChem had not yet been assigned an ITIS TSN. In July 2011, entries in the BioChem taxonomic code table underwent a review during which names were validated: variant spellings were corrected, and common names were associated with valid taxonomic names. This list of revised names was mapped to a list of standardized scientific names contained in the World Register of Marine Species (WoRMS; WoRMS Editorial Board 2013) using their online TaxonMatch tool (www.marinespecies.org/aphia.php?p=match). Each name in WoRMS is associated with a code referred to as the WoRMS AphiaID. The practice of including a standardized code such as the WoRMS AphiaID as an additional field in a species list is recognized as a data management best practice related to the quality control of marine taxa names. Consulting the WoRMS database facilitates the resolution of synonyms and spelling variations inherent in large databases with content from multiple sources. Records in the BioChem taxonomic code table that refer to non-biotic names, such as sand or oil, will obviously not have a corresponding AphiaID. Nevertheless, data associated with these terms are still valuable, and the option exists to output this information along with records referencing valid biota. Even though the taxon code table underwent a major update and correction exercise over the past few years, there remain duplicates and other inconsistencies that will be resolved as time allows.

2.3 BIOCHEM CODE TABLES

BioChem has code tables that precisely define data types, analysis and collection methods, plankton names, and sampling gears among others (Table 1). The basic design of these tables is unique code / element name / element definition. Using standardized, well-defined terms facilitates data exchange and interpretation, and a concerted effort to define terms and share vocabularies within the oceanographic community has been the objective of groups such as the International Oceanographic Data and Information Exchange (IODE) (sites.google.com/site/gebichwiki/vocabularies-2) and SeaDataNet (www.seadatanet.org/Standards-Software/Common-Vocabularies).

The BioChem code tables are generally in a state of flux, with additions of—for example—new plankton taxa, new data types, or new analysis methods. Since these tables are not static and any printed version would be quickly outdated, we provide links to them on the OSD website (<http://www.meds-sdmm.dfo-mpo.gc.ca/biochem/biochem-eng.htm>; see the link "Code Tables" under the heading "Resources"). The tables are available as text files that can be opened with a text editor or in a spreadsheet, which makes sorting simpler. The files will be replaced on the website link when modifications occur, so that the most recent versions should be available.

Table 1. List of files whose contents define codes used in the BioChem database. A link allowing the user to download these files can be found at <http://www.meds-sdmm.dfo-mpo.gc.ca/biochem/biochem-eng.htm> (see the link "Code Tables" under the heading "Resources"). Tables are named "BC_" followed by the table name and the date that the file was created (having the date as part of the filename allows one to know whether information in the table is current). Table content is modified when new codes are needed, so these files will be updated when required to reflect new code additions.

Code table name	Brief description
Tables used for both discrete and plankton data	
BC_Gears_date.txt	Gear used for sample collection; e.g., net, bottle
BC_Preservations_date.txt	How the sample was preserved; e.g., frozen -80°C, buffered 10% formaldehyde
BC_Storages_date.txt	How the sample was stored following collection and preservation; e.g., room temperature, frozen -20°C
BC_Units_date.txt	Variable's unit of measure; e.g., mmol/m ³ , g/kg
BC_Qual_Codes_date.txt	Code indicating the quality of the data or metadata value (largely follows GTSP standard scheme)
Code tables for discrete data	
BC_Data_Types_date.txt	Specific data type defined according to sample handling, preservation, and analysis methods; e.g., filtered sample stored at -80°C and analyzed for chlorophyll <i>a</i> according to the Holm-Hansen method
BC_Analyses_date.txt	Description or reference for the analysis method
BC_Data_Retrievals_date.txt	General category of a variable (data type) that may include many subcategories; e.g., O ₂ includes dissolved oxygen measured by CTD sensor, Winkler titration, bench probe
BC_Sample_Handlings_date.txt	Description or reference giving details of how the sample was handled; e.g., AZMP protocol, sample prefiltered
Code tables for plankton data	
BC_Natnl_Taxon_Classes_date.txt	Hierarchical taxon groupings. The aim is to be able query groups of taxonomically similar organisms
BC_Natnl_Taxon_Codes_date.txt	List of all plankton names with corresponding scientific name authorships and codes (TSN, ITIS_TSN, APHIAID) (see Annex IV)
BC_Procedures_date.txt	Details on the plankton analysis procedure; e.g., quantitative, qualitative, quantitative for select groups only
BC_Collection_Methods_date.txt	Description of different collection methods (usually net deployment details); e.g., vertical, oblique, or horizontal tow (see Annex III)
BC_Volume_Methods_date.txt	Description of how the volume of water sampled was determined; e.g., flow meter, estimated from depth, wire angle, and net mouth opening
BC_Trophic_Descriptors_date.txt	Indicates the organism's trophic position; e.g., autotroph, detritivore
BC_Life_Histories_date.txt	Details included with some plankton identifications indicating the organism's life history stage; e.g., copepod nauplius stage I (note that this table is also used to include sample details that are not related to life history; e.g., damaged and exoskeletons)
BC_Sexes_date.txt	Sex of organism; e.g., male, female, unknown

2.4 QUALITY CONTROL

It is good practice to implement quality control (QC) procedures on data before they are archived. Quality control refers to tests that are applied to data to ensure their accuracy, to identify and correct errors, and to document what quality tests were applied to the dataset. Ideally, quality control should begin with the data collectors, analysts, and principal investigators responsible for the data in question since they are most familiar with the data and would be aware of any events or particularities that might affect data quality. Their comments should be noted and made available to the person who applies the formal QC tests. Reliable, high-quality data are the foundation upon which scientists base accurate conclusions and predictions.

Because a large volume of DFO's biological and chemical data had been without a dedicated database for so long, it was decided that legacy datasets should be loaded to BioChem as soon as possible to assure their safeguard and availability, even though many of these had not undergone a formal quality control exercise. It is hoped that existing errors are few and that these will be brought to the attention of BioChem data managers and addressed as necessary. Ideally, recently acquired data should be quality controlled before being archived in the system. As of the publication of this report, almost half of all the data held in the discrete functional area have been assigned quality flags, although there is currently no mechanism to indicate what—if any—specific tests were applied.

2.4.1 *Quality control of metadata (header information)*

Quality tests examine the accuracy of the metadata (i.e., mission- and event-level information) as well as the data themselves. Tests on the metadata include temporal tests (impossible date/times, whether all events occurred within the mission start–end dates) and spatial tests (impossible positions, positions on land, and positions outside the study area) as well as impossible ship speeds. These tests are applicable to any type of data (i.e., plankton and discrete in our case) and are listed in Table 2.

2.4.2 *Quality control of discrete data*

Quality tests for discrete data include tests for globally and regionally impossible values and tests comparing data values within an event, comparing values to a regional climatology, and comparing all events from the same mission. At present, only Québec Region (IML) has implemented a formal QC procedure for discrete data (Devine and Lafleur 2008), and only the most common data types are currently examined with the defined tests (temperature, salinity, dissolved oxygen, nutrients, chlorophyll; Table 2); other variables are visually checked for reasonable value ranges and outliers. The quality tests applied to discrete data are largely drawn from procedures used by NOAA's National Oceanographic Data Center during the production of the World Ocean Database (Conkright et al. 2002) as well as many of the tests proposed in the GTSP Real-Time Quality Control Manual (Global Temperature–Salinity Profile Program; UNESCO 1990). A complete description of quality tests applied to the metadata and discrete data from Québec Region can be found at the following links: http://slgo.ca/app-sgdo/en/docs_reference/botl_odf_quality.html or in Devine and Lafleur (2008; [11](http://www.meds-</p></div><div data-bbox=)

sdmm.dfo-mpo.gc.ca/isdm-gdsi/azmp-pmza/docs/bulletin_7_05.pdf) (available in English and French).

Climatologies (or atlases) are statistical summaries of in situ measurements. They consist of minimum, maximum, average, and standard deviation values for measurements from a fixed area, time period, and depth range. Climatologies are important components of quality control procedures in that they allow a comparison of the examined data to data already collected (and considered valid) from the same region/season/depth. There are a number of published atlases for temperature and salinity (Petrie et al. 1996 a, b) and nutrients (Petrie et al. 1999, Brickman and Petrie 2003) for Atlantic Canada. When a published atlas is not available, one can be compiled using data from other sources. For example, a rough nutrient climatology was constructed for Newfoundland waters (5×5 degrees) using data from NOAA/NODC's World Ocean Atlas (http://www.nodc.noaa.gov/OC5/WOA09/pr_woa09.html).

Table 2. Quality control tests applied to metadata—which are relevant for both plankton and discrete data—as well as tests performed on discrete bottle data. TEMP: temperature; PSAL: salinity; DOXY: dissolved oxygen; NTRZ: nitrite+nitrate; PHOS: phosphate; SLCA: silicate.

Test name	
GTSP Platform Identification	Metadata
GTSP Impossible Date/Time	Metadata
GTSP Impossible Location	Metadata
GTSP Position on Land	Metadata
GTSP Impossible Speed	Metadata
Cruise Track Visual Inspection	Metadata
Global Impossible Parameter Values	Data
Regional Impossible Parameter Values	Data
Profile Envelope	Data
Constant Profile	Data
Replicate Comparisons	Data
Bottle versus CTD Measurements (TEMP, PSAL, DOXY)	Data
Excessive Gradient or Inversion (TEMP, PSAL, NTRZ, PHOS)	Data
Surface Dissolved Oxygen Data versus Percent Saturation	Data
Petrie Monthly Climatology (TEMP, PSAL)	Data
Brickman Monthly Climatology (NTRA, PHOS, SLCA)	Data
Ratio and Profile Visual Inspection (station data)	Data
Replicates Visual Inspection (whole cruise data)	Data
Bottle versus CTD Measurements Visual Inspection (whole cruise data)	Data
Ratio and Profile Visual Inspection (whole cruise data)	Data
Parameter Patterns With Time (whole cruise data)	Data

2.4.3 Quality control of plankton data

Plankton data are inherently more difficult to quality control since no quantitative tests can be applied to identify unlikely or impossible values. The US agency NOAA/NMFS maintains a global plankton database, the documentation of which contains a discussion on quality control related to plankton datasets (O'Brien 2007; see <http://www.st.nmfs.noaa.gov/plankton/2007/index.html>). O'Brien states that "in general,

plankton data are usually 'correct' in their original source media and any anomalous values found in these data are due to natural processes (e.g., blooms, swarms, patchiness) or mechanical sampling issues (e.g., gear failure or clogged nets)." He adds that errors often occur because of misunderstandings that arise when interpreting datasets from many sources. Most of the plankton datasets archived in BioChem are from DFO regional data centres, so questions can often be resolved by asking the original data collectors.

For current datasets, Québec Region verifies sample depth and filtered volumes by comparing different methods of calculation as well as data resulting from other sensors deployed at the same time. The taxonomic data are validated by looking for duplicates. If the original data compilation sheet is available, split and dilution information are verified and the spellings of taxonomic names are checked.

Plankton datasets should be reviewed prior to archive to check for inclusion of zeros or negative or null values in the data fields. It is also important to note whether missing value indicators were included in the source dataset. Most plankton datasets in BioChem do not include zeros in the counts field. Even though there may be no counts associated with a specific taxonomic name on a sample analysis sheet, it does not necessarily follow that the taxon was not present. It is possible that it was included in a group or was assigned to a higher taxonomic rank. For example, one dataset may have counts for *Pseudocalanus*, another dataset may have combined this genus into a group called *Pseudocalanus* / *Clausocalanus* / *Paracalanus*, and yet another might have included it in a group called juvenile copepods (Copepoda copepodite stages I-V).

Plankton datasets result from the partial (only certain groups are targeted and counted) or the complete identification of all organisms found in the sample. In addition, datasets often include a mixture of quantitative and qualitative data. Groups of primary interest may be identified to the species level and perhaps even further—to life history stage and sex—while other taxa may be identified only to a higher taxonomic ranking such as genus, family, or class; yet others may be noted as present/absent or rare/abundant. For instances where only qualitative information is available, the taxonomic name is entered, the presence flag is selected, and a note may be added in the collectors comment field. It is important to record these distinctions in the analysis method and archive them with the data to ensure correct interpretation. The relevant code table associated with this info is BCProcedures.

Most zooplankton data are abundances recorded as counts of individuals found in a sample (or subsample); however, some legacy datasets do not include raw data values but have been transformed to standardized units such as numbers per cubic metre or per square metre. Similarly, phytoplankton are often reported as numbers of cells per litre and bacteria as cells per millilitre. In all these cases, counts are quantitative; one must know the volume of water sampled (filtered by the deployed sampling gear or subsampled from a water bottle), the depth strata sampled (beginning and end), and the dilution of the sample (the split fraction) in order to calculate the number of individuals per cubic metre or per square metre. If one of these variables is missing, the data are considered as qualitative and cannot be used to calculate standardized abundances.

2.4.4 Quality control flags

The quality flags assigned to data held in BioChem largely follow the flag meanings as defined by GTSP (UNESCO 1990) (Table 3). The exception is QC=8, which is not assigned in the GTSP scheme. We use this flag to indicate that data were examined by an authority and pronounced correct; however, there is no record of whether or what formal tests were applied to these data.

Table 3. Quality control flags used in BioChem and their meanings. The quality flags of higher value take precedence over those with lower value; e.g., a QC flag of 0 has a lower priority than a flag of 9. As such, if a test judged a data value as doubtful (flag 3) and the following test judged it as erroneous (flag 4), the quality flag 4 would be retained.

Flag	Meaning
0	no quality control
1	value seems correct
2	value appears inconsistent with other values
3	value seems doubtful
4	value seems erroneous
5	value was modified; now seems correct
6, 7	unassigned
8*	<i>QC was performed by data producer</i>
9	value missing

* not a standard GTSP quality flag

3.0 USING BIOCHEM

BioChem's home webpage is hosted by OSD (<http://www.meds-sdmm.dfo-mpo.gc.ca/biochem/biochem-eng.htm>). This page includes a description of the project, information on how to register as a user, citation recommendations, and links to documentation related to the database. This page also contains links to applications that allow users to query the database and data managers to archive data.

Anyone can use BioChem after first obtaining a username and password using the registration form on the website. Users are asked to provide email contact information and organization affiliation. New users are assigned to one of the three levels of BioChem users—user, data manager, or super user. Each level has different privileges: a user is only able to query and view existing records and cannot modify any records; a data manager can query, insert new records, and edit existing records via application forms; and a super user has the all of the privileges of a data manager and is also able to create BioChem user accounts and modify code tables.

Users access data archived in BioChem through the query application, which queries either the discrete or the plankton functional area (see section 3.1). Criteria are specified to identify the data of interest and/or to limit the size of the query return; these can include items such as

geographic area, date, and data type. A user manual is available from the query webpage that provides step-by-step instructions for building and submitting a query.

Data managers from DFO regional data centres load data to the archive using the BioChem edit application. Data are initially loaded into BioChem station and BioChem data tables (see section 3.2) in temporary data manager accounts, where they are validated according to rules that were specified during the database design. Once validated, the data are transferred from the data manager account to the database.

3.1 RETRIEVING DATA USING THE QUERY APPLICATION

The BioChem query application is used to build queries. Submitted BioChem queries are run off line, and the user receives an email (usually within a few minutes) when the results are ready. This email gives a link to the DFO FTP site where the resulting compressed ("ZIP") file can be retrieved. The file remains available for seven days. Note that although the application and the email notification are bilingual, the query results as well as code table terms and definitions are in English only.

The ZIP file package contains a number of files that include the query specifications and the data extracted based on these specifications. Most files are common to both the discrete and plankton functional areas (Table 4). The BioChem_Query_####_Data.csv file contains the data. This file looks very different depending on whether the query was directed to the discrete or to the plankton functional area. Guidelines on how to interpret this file are contained in Annex V.

Table 4. Description of files in query return ZIP file.

File name	Description of file contents
BioChem_Query_####.txt	Contains a record of the query criteria
BioChem_Query_####_Collectors.csv	Lists the collectors associated with the returned data; the "collector" can be the name of a person, a project, a program, a published report, or other. This identifier may be used to link to additional metadata in the DFO-MEST where it is referred to as the BioChem series identifier
BioChem_Query_####_Counts.csv	Gives the number of records associated with the individual data types or taxon codes returned in the query. For discrete data, the units and data format are also given
BioChem_Query_####_Data.csv	Contains the data and related metadata corresponding to the query specifications (see Appendix V for details on how to interpret this file)
BioChem_Query_####_Locations.csv	Contains a list of the sampling events returned by the query (mission name, event ID, station name, lat/long, start date/time) (This is a subset of the information in the *_Data.csv file)
BioChem_Query_####_Missions.csv	Lists all the missions considered during the query (default is all missions in the database when the query was run). Missions are identified by their unique descriptor
BioChem_Query_####_Samples.csv (Plankton queries only)	Contains a list of the events and associated metadata of records included in the *_Data.csv file. (This is a subset of the information in the *_Data.csv file)

3.1.1 Citing data from BioChem

BioChem users are requested to cite the database and records that they downloaded as follows:

DFO (*insert year database accessed here*). BioChem: database of biological and chemical oceanographic data. Department of Fisheries and Oceans, Canada.
<http://isdms.gc.ca/biochem/biochem-eng.htm>. Database accessed on *insert date here*

In addition, this report should be included in the reference section:

Devine, L., M.K. Kennedy, I. St-Pierre, C. Lafleur, M. Ouellet, and S. Bond. 2014. BioChem: the Fisheries and Oceans Canada database for biological and chemical data. Can. Tech. Rep. Fish. Aquat. Sci. 3073: v + 40 pp.

3.2 LOADING DATA USING THE EDIT APPLICATION

Data are loaded to the BioChem database by data managers from the various DFO regions. The procedures used to archive discrete and plankton data are similar: source data are used to populate two tables referred to as the BioChem station table and the BioChem data table (Annex VI). A series of documents were created during database development (~2002–2006) to record descriptions of the database; these included data dictionaries. Although these documents refer to an older version of BioChem, much of the information is still relevant (see the link under “Resources” at <http://www.meds-sdmm.dfo-mpo.gc.ca/biochem/biochem-eng.htm>). The definitions of items required to populate the load tables are from these documents, although it has been necessary to create a few new fields to accommodate unforeseen situations.

Once these tables have been created, they should be checked over carefully to avoid problems during loading. For example, one must make sure that dates and positions are in the correct formats, that data values fall within the allowed minimum and maximum values for a given variable, and that all required fields are populated. The load tables are transferred to temporary tables in the database via SQL*loader, insert SQL command, or any other database application program that has a tool to load or import data. The system applies validation checks to these temporary tables before final loading to the database. If a validation test fails, it is often difficult to identify the source of the error, and it may be best to abort the load and examine the dataset for errors before proceeding. When data are successfully loaded, `CREATED_DATE` and `CREATED_BY` (to identify the user ID of the person loading the data) are added to the records.

4.0 BIOCHEM AND GLOBAL DATA INITIATIVES

DFO's data management policy states that data must be secure, discoverable, and accessible; BioChem fulfills these objectives for bottle and plankton data. In addition, data stored in BioChem are standardized and terms can be mapped to recognized vocabularies so that collaboration and exchange are facilitated. Such collaborations include submitting data to global databases such as the US-NODC's World Ocean Database (WOD, http://www.nodc.noaa.gov/OC5/WOD/pr_wod.html) and data exchange with international initiatives such as the Carbon Dioxide Information Analysis Center (CDIAC,

<http://cdiac.ornl.gov/oceans>) and the Ocean Biogeographic Information System (OBIS, <http://www.iobis.org/>). Scientific names (usually older synonyms from legacy datasets) are submitted to WoRMS for inclusion in their database. Data flow paths from the collection (regional) level to global (international) collaborations are shown in Figure 6.

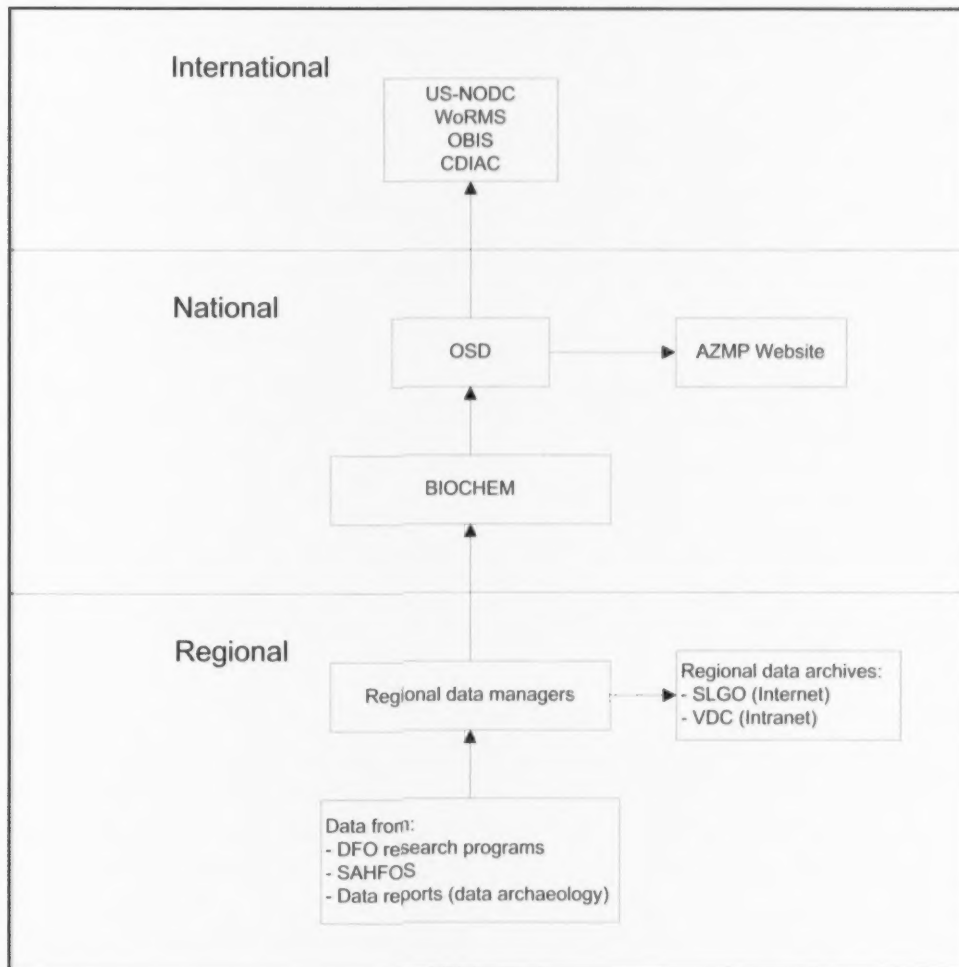


Figure 6. Schematic of data flow paths from data collection through international exchanges.

5.0 ACKNOWLEDGEMENTS

We acknowledge the efforts of many individuals—too numerous to name here—who participated in the early phase of database design, worked to get the database online and functional, loaded data, and provided comments to improve the database's functionality. Notable among these were Savi Narayanan (Ottawa), Doug Gregory (BIO), and John O'Neill (BIO), who played important roles in the early stages of the database's inception; Claude Guay (Ottawa), who acted as steering committee chair from 2006–2010; Pierre Clement, Peter Strain, Robert Benjamin, and Tony Joyce (BIO), who loaded much of the BIO legacy data; and Greg Levonian (Ottawa), who worked hard to resolve informatics problems. Assistance and comments from Bernard Pelchat (IML), Brian Petrie (BIO), and Jackie Spry (Sprytech Biological Services) were much appreciated. Pierre Clement and Dave Senciall (NAFC) offered constructive comments during the final review of this report.

6.0 REFERENCES

- Brickman, D., and B. Petrie. 2003. Nitrate, silicate and phosphate atlas for the Gulf of St. Lawrence. Can. Tech. Rep. Hydrogr. Ocean Sci. 230: xi+152 pp. <http://www.dfo-mpo.gc.ca/Library/274861.pdf>
- Canadian Oceanographic Data Centre. 1972. CODC: A decade of growth, 1962-1972. 26 pp. CATNO 50987.
- Conkright, M.E., J.I. Antonov, O. Baranova, T. P. Boyer, H.E. Garcia, R. Gelfeld, D. Johnson, R.A. Locarnini, P.P. Murphy, T.D. O'Brien, I. Smolyar, and C. Stephens. 2002. *World Ocean Database 2001*, Volume 1: Introduction. Ed: Sydney Levitus, NOAA Atlas NESDIS 42, U.S. Government Printing Office, Washington, D.C., 167 pp.
- Devine, L., and C. Lafleur. 2008. Quality control of bottle data at Maurice Lamontagne Institute. AZMP Bulletin PMZA 7: 27–37. http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/azmp-pmza/docs/bulletin_7_05.pdf
- Gregory, D., and S. Narayanan. 2003. BioChem: A national archive for marine biology and chemistry data. AZMP Bulletin PMZA 3:11–13. http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/azmp-pmza/docs/bulletin_3_04.pdf
- Kennedy, M., and L. Bajona. 2009. A data manager's guide to marine taxonomic code lists. Can. Tech. Rep. Fish. Aquat. Sci. 2827: iii + 23 pp. <http://www.dfo-mpo.gc.ca/Library/336700.pdf>
- O'Brien, T.D. 2007. COPEPOD: The Global Plankton Database. A review of the 2007 database contents and new quality control methodology. U.S. Dep. Commerce, NOAA Tech. Memo., NMFS-F/ST-34, 28 pp.
- Petrie, B., K. Drinkwater, D. Gregory, R. Pettipas, and A. Sandström. 1996a. Temperature and salinity atlas for the Scotian Shelf and the Gulf of Maine. Can. Tech. Rep. Hydrogr. Ocean Sci., 171: v+398 pp. <http://www.dfo-mpo.gc.ca/Library/193505.pdf>

- Petrie, B., K. Drinkwater, A. Sandström, R. Pettipas, D. Gregory, D. Gilbert, and P. Sekhon. 1996b. Temperature, salinity and sigma-t atlas for the Gulf of St. Lawrence, Can. Tech. Rep. Hydrogr. Ocean Sci., 178: v+256 pp. <http://www.dfo-mpo.gc.ca/Library/198079.pdf>
- Petrie, B., P. Yeats, and P. Strain. 1999. Nitrate, silicate and phosphate atlas for the Scotian Shelf and Gulf of Maine. Can. Tech. Rep. Hydrogr. Ocean Sci. 203: vii+96 pp. <http://www.dfo-mpo.gc.ca/Library/238043.pdf>
- UNESCO. 1990. GTSP real-time quality control manual. Intergovernmental Oceanographic Commission, Manuals and Guides no. 22.
- WoRMS Editorial Board. 2013. World Register of Marine Species. Available from <http://www.marinespecies.org> at vliz. Accessed 2013-12-10.

ANNEX I. LIST OF ACRONYMS AND ABBREVIATIONS USED IN THIS DOCUMENT

Abbreviation	Meaning
AphiaID	Unique taxon code used by the World Register of Marine Species
AZMP	Atlantic Zone Monitoring Program (DFO)
BIO	Bedford Institute of Oceanography (DFO, Dartmouth, NS)
BioChem	DFO's database of biological and chemical oceanographic data
BODC	British Oceanographic Data Centre
CCIW	Canadian Centre for Inland Waters (DFO, Burlington, ON)
CDIAC	Carbon Dioxide Information Analysis Center
CODC	Canadian Oceanographic Data Centre (formerly a DEMR, Marine Sciences Branch, division, became MEDS in 1973)
CPR	Continuous plankton recorder
CTD	Conductivity, temperature, depth probe
DEMR	Department of Energy, Mines and Resources (now in Natural Resources Canada)
DFO	Fisheries and Oceans Canada
DFO MEST	DFO metadata entry and search tool
ERD	Entity relationship diagram
FTP	File transfer protocol
GFC	Gulf Fisheries Centre (DFO, Moncton, NB)
GTSP	Global Temperature-Salinity Profile Program
HPLC	High performance liquid chromatography
ICES	International Council for the Exploration of the Seas
IML	Institut Maurice-Lamontagne (DFO, Mont-Joli, QC)
IOC	Intergovernmental Oceanographic Commission
IODE	International Ocean Data and Information Exchange (program of the IOC)
IOS	Institute of Ocean Sciences (DFO, Sidney, BC)
ISDM	Integrated Science Data Management (DFO, Ottawa, ON), formerly MEDS
ISO	International Organization for Standardization
ITIS	Integrated Taxonomic Information System
JCOMM	Joint Technical Commission for Oceanography and Marine Meteorology
JGOFS	Joint Global Ocean Flux Studies
MEDS	Marine Environmental Data Service (DFO, Ottawa, ON), formerly CODC
MONC	Moncton (also called GFC; DFO, Moncton, NB)
NAFC	North Atlantic Fisheries Centre (DFO, St John's, NL)
NMFS	National Marine Fisheries Service (US)
NOAA	National Oceanic and Atmospheric Administration (US)
NODC	National Oceanographic Data Center (US)
OBIS	Ocean Biogeographic Information System
OSD	Oceanography and Scientific Data (DFO, Ottawa, ON), formerly ISDM
QC	Quality control
SABS	St. Andrews Biological Station (DFO, St. Andrews, NB)
SAHFOS	Sir Alister Hardy Foundation for Ocean Science (UK)

SLGO	St. Lawrence Global Observatory (web portal)
TSN	Taxonomic Serial Number
VDC	Virtual Data Centre (BIO intranet)
WOD	World Ocean Database (suffix numbers indicate release year, example WOD94) (US NODC)
WoRMS	World Register of Marine Species
Y2K	Year 2000

ANNEX II. MISSION DESCRIPTORS

The Oceanography and Scientific Data branch (OSD, DFO, Ottawa) assigns mission descriptors to identify missions; these descriptors are used to identify datasets held in BioChem. The primary function of the mission descriptor is to identify the data collection so that it is easy and unambiguous to recognize and so that tracking its history is simplified. This annex provides details on how these descriptors were and are assigned. We present here a detailed account of these codes—even beyond the scope of those used in BioChem—with the aim of documenting the evolution of these codes through time.

Two conventions co-exist in the OSD archive. In the older convention, the mission descriptor was a nine-digit code constructed as follows:

01–02 Country, according to IOC country codes (<http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/code/list/004-eng.asp>) (Canada is 18)

03–04 Institution, according to OSD (column "Old Code"; <http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/code/list/010-eng.asp>)

05–06 Last two digits of the year when the first station was sampled

07–09 Original mission identification number with leading zeros.

This scheme, without the country code prefix, was known as the CODC (Canadian Oceanographic Data Centre) scheme from 1962 to 1972. Characters 03–04 made the descriptor unique nationally, and characters 01–02 made it unique internationally.

In the more recent convention, OSD (then MEDS) adopted the following mission descriptor scheme:

01–04 ICES Platform code

05–06 Last two digits of the year when the first station was sampled

07–09 Numerical portion of the original mission identification number with leading zeros. If a mission number was not provided, then a unique identifier is assigned by OSD.

Examples:

- 18VAyy669 is the mission descriptor for monthly sampling at Prince 5, where yy is the last two digits of the observation year and VA means that small boats were used for sampling
- 18HU08003 identifies data collected on the CCGS *Hudson* in 2008 with the original regional mission number of HUD2008003
- 18HE12004 identifies data collected by helicopter in 2012; the original regional mission number was 2012004.

Special platform codes are used in some circumstances. One commonly used is 18VA; this is for cases when

- the platform is known but does not have its own code (usually because the vessel is too small)

- more than one platform was used during the mission (sometimes coded 1890)
- the mission occurred over an extended period of time with the vessel returning to shore between sampling activities (more than one platform may have been used; sometimes coded 1890).

When the platform is unknown, 1899 is used. Sampling done on foot, e.g., from a wharf or bridge, is coded 187F. There are some obsolete and nonstandard mission codes in the database, and—over the years and depending on the personnel involved—different codes have occasionally been used for the same situation (e.g., 18VA and 1890 as described above). This is unfortunate but reveals another challenge of coordinating data flow among different groups. A list of platform codes can be found at the NOAA / NODC website (<http://www.nodc.noaa.gov/General/NODC-Archive/platformlist.txt>), but the authoritative list is with ICES and SeaDataNet (<http://vocab.ices.dk/>; [http://seadatanet.maris2.nl/v_bodc_vocab/search.asp?name=\(C174\)%20SeaDataNet+Cruise+Summary+Report+ship+metadata&l=C174](http://seadatanet.maris2.nl/v_bodc_vocab/search.asp?name=(C174)%20SeaDataNet+Cruise+Summary+Report+ship+metadata&l=C174)), known as the controlled vocabulary list C174. It references the BODC/SeaDataNet L06 Platform Categories, which sorts platforms into groups such as ships, helicopters, ice islands, and so on (http://seadatanet.maris2.nl/v_bodc_vocab/welcome.aspx/). SeaDataNet has a pending project to remove all of the *90, *99, and all other "catch-all" codes like 18VA from the platform code list.


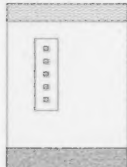
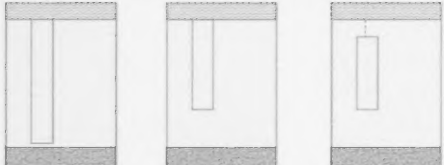

It is unknown when the newer convention came about, but it was already used in 1990 (the first year for which we can trace the database's history), though not predominantly. Use of both conventions endured until 2005, with the attribution of mission descriptors in the newer convention becoming predominant in 1996.






Older mission codes sometimes used the country code followed by 00, or " " (double blank), when the platform was unknown or did not have a code. Canadian missions collecting bathythermograph data between 1943 and 1965 mostly have mission descriptors starting with 1800. Until 1972, the CODC had separate databases for bottle and bathythermograph data (CODC 1972); it is likely that the origin information of pre-1966 bathythermograph data was not preserved in the data form when OCEANS V, the first CODC system allowing storage of bathythermograph and bottle data in a common data format, was created. This is the archive that MEDS (then ISDM, then OSD) inherited and built upon.

The problem with most codes is that once adopted they tend to become obsolete. This is the case for the two-digit country codes, which constitute the first two digits of the platform codes. IODE/JCOMM recently recommended the ISO 3166 standard for identifying countries in oceanographic data exchange. There is no plan to upgrade the mission descriptor scheme to use ISO 3166. A caveat with the current scheme is that there is no easy way to sort in chronological order: oceanographic observations span more than a century already and only two digits are used for the year. By the same token, mission descriptors used for data collected 1914 cannot be reused in 2014. This should not be a problem because most ships have a lifespan of less than 100 years, and the platform names are not recycled, but it is problematic for cruises with generic (or multiple-hull) prefixes such as 187F, 18VA, 1890, 18HE, and 1899.

ANNEX III. PLANKTON COLLECTION METHODS

While most of BioChem's code tables refer to terms that are commonly used in the oceanographic community, the plankton collection method table uses terms that may be unfamiliar. To adequately document plankton sampling, we not only need to note the type, model, and size of the sampling gear, but also record how the device was deployed. For example, it should be recorded whether the gear was towed horizontally, vertically, or obliquely, or if the sample was collected from a bottle sampler such as a Niskin or from an in situ pump. Additional information may be required; for example, when the sample was collected using a multiple net device (e.g., BIONESS, MOCNESS), we need to know whether the sample is from one net or the whole cast (i.e., if samples were pooled). The table below illustrates a subset of the collection methods currently available in BioChem. This list will be extended as needed, so it is best to consult the BC_Collection_Methods code table for the most recent list.

Term	Definition / Example	
Hydrographic	Discrete sample collected from one depth. Examples: bottle closed at 10 m; pump lowered to 10 m before sample collected.	
Hydrographic – integrated	Samples collected from discrete depths were pooled to form an integrated depth sample. Example: water samples from 0, 15, 25, 50, 100 m pooled for an integrated 0–100 m sample.	
Vertical	Sample collected between two depths from a stationary platform. Examples: plankton tow from bottom–surface, from mid–water to surface, or opening–closing net sampling 100 m to 50 m.	
Horizontal	Sample collected within a defined depth stratum from a moving platform for a given period of time or distance (often at a given speed).	

Surface	Sample collected horizontally at the surface from moving platform for a given period of time or distance (gear mouth opening may be partially above water surface). Example: net towed at air/sea interface for 15 min.	
Oblique	Sample collected between two depths from a moving platform. Gear is opened at one depth and closed at a second; rate of gear retrieval is kept relatively constant.	
Double oblique	Sample collected between two depths from a moving platform. Gear is open at the surface and fishes down to specified depth and then fishes back up to the surface.	
Stepped oblique	Sample collected between two depths from a moving platform. Gear is opened at one depth and closed at a second depth. The rate of gear retrieval is not constant, allowing it to fish specific depth strata before being raised to fish another stratum.	
Yoyo or sawtooth oblique	Sample collected within a specific depth stratum from a moving platform.	

ANNEX IV. BIOCHEM TAXONOMIC CODE TABLE

BioChem's taxonomic code table is certainly the largest in the database. It contains the 11 fields noted in the list below. The meanings of these terms and a discussion of the different species codes are the object of this annex. Those items marked with → will be covered in greater detail in the text below the list.

- NATIONAL_TAXONOMIC_SEQ: the system-generated number associated with a given name.
- DATA_CENTER_CODE: the data centre that added a specific code to the database (90: general, 10: NAFC, 20: BIO, 25: GFC, 30: IML, 40: OSD, 50: IOS, 60: Central & Arctic [CCIW]).
- TAXONOMIC_NAME: name assigned to the identified organism, grouping of organisms, or item.
- AUTHORITY: the name of the person credited with describing the organism and often the year of publication. This is commonly referred to as the authorship. If the authorship is in parentheses, this means that the organism was reclassified and the genus/species name is not the one appearing in the original publication.
- COLLECTORS_COMMENT: this field sometimes contains a reference to the identification guide or publication used for specimen identification, but it may contain any other comments that the data manager who adds the code table record feels is pertinent.
- DATA_MANAGERS_COMMENT: this field often gives the name of the program under which the data was collected that resulted in the name being added to the code table, but it may contain any other comment that the data manager who adds the code table record feels is pertinent.
- SHORT_NAME: the taxon's shortened nickname (used for some datasets but not generally populated).
- BEST_NODC7: a modified version of the NODC taxonomic code.
- TSN_ITIS: the unique code used by the Integrated Taxonomic Information System (<http://www.itis.gov/>)
- TSN: the taxonomic serial number assigned for use in BioChem.
- APHIAID: the unique code used by the World Register of Marine Species (<http://www.marinespecies.org/>).

One of the first major attempts to assign species codes was led by the US NODC (<http://www.nodc.noaa.gov/General/CDR-detdesc/taxonomic-v8.html>). Its system (up to version 7) kept the notion of taxonomic hierarchy in the codes. They were known as "smart" codes because examination of the first few digits allowed one to determine the kingdom, phylum, and so on of the organism in question. This "smart" coding system had to be abandoned when the system ran out of numbers within certain taxa. BEST_NODC7 was included in the original BioChem design with the idea of facilitating sorting by taxonomic hierarchy. However, this method of grouping taxa has become outdated and the field is often not populated. When the NODC taxonomic code system was abandoned, that organization joined with other US and

international groups to form ITIS (now known as the Integrated Taxonomic Information System; <http://www.itis.gov/>). Its goal is to create an easily accessible database with reliable information on scientific names and their hierarchical classification. ITIS assigns a TSN—taxonomic serial number—to taxa.

While ITIS is a robust coding system, its emphasis is not marine, thus organisms encountered in plankton datasets are frequently not yet included. For this reason, BioChem data managers looked to the World Register of Marine Species (<http://www.marinespecies.org/>). The aim of WoRMS "...is to provide an authoritative and comprehensive list of names of marine organisms, including information on synonymy. While highest priority goes to valid names, other names in use are included so that this register can serve as a guide to interpret taxonomic literature." WoRMS has a consolidated database called Aphia, which "...contains valid species names, synonyms and vernacular names, and extra information such as literature and biogeographic data. Besides species names, Aphia also contains the higher classification in which each scientific name is linked to its parent taxon. The classification used is a 'compromise' between established systems and recent changes. Its aim is to aid data management, rather than suggest any taxonomic or phylogenetic opinion on species relationships" (WoRMS Editorial Board 2013).

Taxonomic names in a dataset are not always straightforward. Records in taxonomic lists may include entries other than scientific names. For example, they can be groups of valid species, common names, and non-living material as well as unverified/questionable names. All entries in BioChem are assigned a TSN; when possible, it is the same as the TSN_{ITIS}. However, some valid groups have not yet been assigned a TSN by ITIS or an AphiaID code by WoRMS, and some entries (e.g., taxon groups or non-living material) will never have one of these. Nevertheless, a code must still be assigned. BioChem data managers use the same convention developed for use in NODC's World Ocean Database (WOD). This scheme assigns negative codes to items falling in this category. In a sense, these negative codes are "smart" codes, since their value indicates why they do not yet have a valid ITIS TSN. The list below provides an explanation of how negative TSNs are assigned.

–9*: These are for name descriptions that are non-living items. Example: sand, rocks, plastic, garbage, metal, glass, bone, wood.

–7*: These are for taxon names that were still under ITIS review at the creation of WOD01. Upon completion of the review, most will be assigned an official ITIS TSN that will replace the temporary –7 code. Those that fail to meet the ITIS review criteria will be reassigned a –1 code value.

–6*: These are for taxon names that contain two or more taxonomic groups, making it difficult to assign a single ITIS taxonomic code that preserves the original meaning. Example: "salps and doliolids"; both are legitimate by themselves, but combined they cannot be matched to a single ITIS TSN.

–5*: These are for groupings that are not taxonomic or that cover too many taxonomic groups to assign a single ITIS taxonomic code that preserves the original meaning. Example: shrimp, worms, plankton.

–1*: These are for taxon names that were submitted to ITIS and failed to meet the ITIS review criteria. The descriptions may be not taxonomic, non-existent, or misspelled beyond recognition.

ANNEX V. HOW TO INTERPRET BIOCHEM QUERY OUTPUT

The BioChem query application extracts data fitting the specified criteria and creates a ZIP file that is sent to an FTP server. The user is sent an email and must retrieve the results within seven days. Section 3.1 describes the different files contained in the query return. The interpretation of most of these files is clear, with the exception of the most important file—that which includes the data. The lack of clarity in this file is mostly related to the cryptic nature of the codes used for the different types of data. Here we describe the fields in the BioChem_Query_*_Data.csv file, but the code tables must be consulted for full descriptions. These code tables are too voluminous to reproduce here; in addition, codes are regularly added, so any printed list would be quickly outdated. For these reasons, we give the following link where the code tables are available and updated on a regular basis: <http://www.meds-sdmm.dfo-mpo.gc.ca/biochem/biochem-eng.htm> (see the link “Code Tables” under the heading “Resources”).

(a) Example data file from a query on the discrete functional area, BioChem_Query_*_Data.csv

Column name	Example content	Comment
QRY_NO	1179	Application-generated query number
DATA_CENTER	IML	Data centre responsible for collecting and/or archiving the dataset
DESCRIPTOR-NAME	18HU12044-IML-2012-44 Ice/PMZA 2012	Concatenated field: OSD-assigned mission descriptor + mission name
AREA_NAME	Anticosti	Area queried (chosen by user during query)
COLLECTOR_EVENT_ID	IML2012044501	Event ID assigned by collector or data manager
COLLECTOR_STATION_NAME	501	Station ID assigned by collector or data manager
START_LAT	49.29867	Latitude at beginning of sampling (N positive)
START_LON	-64.69333	Longitude at beginning of sampling (E positive)
START_DATE	26-Oct-2012	Date at beginning of sampling (UTC)
START_TIME	2135	Time at beginning of sampling (UTC)
SOUNDING	377	Water depth (m)
COLLECTOR_DEPLOYMENT_ID	IML2012044501	Deployment ID (may be the same as event ID unless multiple deployments during a same event; may be null)
COLLECTOR_SAMPLE_ID	347862	Sample ID
START_DEPTH	359.69	Start depth (m)
END_DEPTH	359.69	End depth (m) (generally = start for discrete data)
COLLECTOR (RESPONSIBLE_GROUP)	AZMP_VILLENEUVE,FRANCOIS (DAISS)	Person responsible for the data; the prefix (if present) may indicate the program under which data was collected

Column name	Example content	Comment
GEAR_MODEL (GEAR_SIZE)	multiple	See code table BC_Gears
Salinity_datatype	Salinity_Sal_PSS_Cal ¹	See code table BC_Data_Types
Salinity	34.96	Data value ²
Salinity_qcCode	1	See code table BC_Qual_Codes
Temperature_datatype	Temp_CTD_1990 ¹	See code table BC_Data_Types
Temperature	5.59	Data value ²
Temperature_qcCode	1	See code table BC_Qual_Codes
O2_datatype	O2_Winkler_Auto ¹	See code table BC_Data_Types
O2	121.22957	Data value ²
O2_qcCode	1	See code table BC_Qual_Codes
Silicate_datatype	SiO4_Tech_SF ¹	See code table BC_Data_Types
Silicate	40.415	Data value ²
Silicate_qcCode	1	See code table BC_Qual_Codes
Phosphate_datatype	PO4_Alp_SF ¹	See code table BC_Data_Types
Phosphate	1.86	Data value ²
Phosphate_qcCode	1	See code table BC_Qual_Codes
Nitrate_datatype	NO2NO3_Alp_SF ¹	See code table BC_Data_Types
Nitrate	24.8	Data value ²
Nitrate_qcCode	1	See code table BC_Qual_Codes
Chlorophyll_A_datatype	Chl_a_Welschmeyer_sf ¹	See code table BC_Data_Types
Chlorophyll_A	1.23	Data value ²
Chlorophyll_A_qcCode	1	See code table BC_Qual_Codes
CHLSENSOR_invivo_datatype	Chl_a_CTD_Fluor ¹	See code table BC_Data_Types
CHLSENSOR_invivo	0.026	Data value ²
CHLSENSOR_invivo_qcCode	0	See code table BC_Qual_Codes

1: Data type name is often a concatenation of variable-name_analysis-method_modifier

2: Units for data values are found in file BioChem_Query_*_Counts.csv

(b) Example data file from a query on the plankton functional area, BioChem_Query_*_Data.csv

Column name	Example content	Comment
PLANKTON_SEQ	20000000151329	Auto-generated unique sequence identifier
QRY_NO	1180	Application-generated query number
DATA_CENTER	IML	Data centre responsible for collecting and/or archiving the dataset
DESCRIPTOR-NAME	18HU12044-IML-2012-44 Monito sections PMZA/Iceforecast	Concatenated field: OSD-assigned mission descriptor + mission name
AREA_NAME	Anticosti	Area queried (chosen by user during query)
COLLECTOR_EVENT_ID	IML2012044TASO1_66	Event ID assigned by collector or data manager
COLLECTOR_STATION_NAME	TASO1	Station ID assigned by collector or data manager
START_LAT	49.216	Latitude at beginning of sampling (N positive)
START_LON	-64.80733	Longitude at beginning of sampling (E positive)
START_DATE	30-Oct-2012	Date at beginning of sampling (UTC)
START_TIME	433	Time at beginning of sampling (UTC)
SOUNDING	240	Water depth at beginning of sampling (m)
COLLECTOR_SAMPLE_ID	IML1244-066	Sample ID
COLLECTOR_DEPLOYMENT_ID	TASO1166202FOND- SURFACEF	Deployment ID
START_DEPTH	235.7	Collection start depth (m)
END_DEPTH	0	Collection end depth (m)
COLLECTOR (RESPONSIBLE_GROUP)	AZMP_PLOURDE,STEPHANE (IML)	Person responsible for the data; the prefix (if present) may indicate the program under which data was collected
GEAR_MODEL (GEAR_SIZE)	Ring net 0.75m (0.75 m)	See code table BC_Gears
MESH_SIZE	202	Net mesh size (µm)
VOLUME	104.173	Volume of water filtered (m ³)
PLANKTON_GENERAL_SEQ	20000002019048	Auto-generated unique sequence identifier
BEST_NODC7	6118010200	See code table BC_Natnl_Taxon_Codes (Annex IV)
TSN	-6902	BioChem-assigned taxonomic serial number (identical to TSN_ITIS, if available); see code table BC_Natnl_Taxon_Codes (also Annex IV)
TAXONOMIC_NAME	Calanus finmarchicus/C. glacialis	See code table BC_Natnl_Taxon_Codes
STAGE_MOLT_NUMBER	copepoditeV	See code table BC_Life_Histories

Column name	Example content	Comment
SEX	UNASSIGNED	See code table BC_Sexes
TROPHIC_DESCRIPTOR	unassigned	See code table BC_Trophic_Descriptors
SOURCE	COLLECTION DE DONNEES	Name of the source of the data (e.g., individual, data collection)
MIN_SIEVE		Often net mesh size, but if sample was size fractionated, minimum sieve size (this had been used in the past to indicate min size, e.g., Copepoda egg (0.135-0.165 mm))
MAX_SIEVE		If sample was size fractionated, maximum sieve size (this had been used in the past to indicate max size, e.g., Copepoda egg (0.135-0.165 mm))
C2	45.2516	Count per m ² (integrated water column)
C3	0.192	Count per m ³
WW2	1.5634	Wet weight per m ² (integrated water column wet biomass)
WW3	0.0066	Wet weight per m ³ (wet biomass)
DW2	0	Dry weight per m ² (integrated water column dry biomass)
DW3	0	Dry weight per m ³ (dry biomass)
V2	0	Biovolume per m ² (integrated water column, settled volume)
V3	0	Biovolume per m ³ (settled volume)
PRESERVATION	formaldehyde, buffered, 4 %	See code table BC_Preservations
METERS_SQD_FLAG	Y	Y/N whether the sample may be used for water column integration
SPLIT_FRACTION	1	Indicates the proportion of the sample analyzed (e.g., 1=whole sample, 0.5=sample divided in two, 0= sample not analyzed)

ANNEX VI. BIOCHEM LOAD TABLES

Below are descriptions of the BioChem station and data tables used to load discrete and plankton datasets. The shaded entries are auto-generated. Note that there is intentional redundancy between the two load tables: these fields must be identical and are used to make the link between the two tables. It is important to respect the rules of the "Column data type" and "Null" columns to avoid generating errors during data loading.

(a) BCDISCRETESTATNEDITS (BCS): station information for discrete data

Column name	Column data type	Null	Column comment	Best practices: examples (if applicable)
DIS_SAMPLE_KEY_VALUE	VARCHAR2(50)	N	Generated sequence number providing unique numeric values for the BCDISCRETESTATNEDITS primary key column	
MISSION_DESCRIPTOR	VARCHAR2(50)	N	Cruise number (or similar)	Code assigned by OSD, ensures national coordination; e.g., 18IU12044 (see Annex II)
EVENT_COLLECTOR_EVENT_ID	VARCHAR2(50)	N	Number assigned to station during data collection	Unique identifier for the sampling event, may be mission# + consecutive//; e.g., IMI.201204411 (mission IMI.2012-44, first station, first cast)
EVENT_COLLECTOR_STN_NAME	VARCHAR2(50)	Y	Descriptive name of station	Can be descriptive station or location name, or consec//, e.g., 11 (first station, first cast at that station)
MISSION_NAME	VARCHAR2(50)	Y	Mission name	Originator's mission number and/or common name(s) for the mission; e.g., IMI.2012-44 Ice/PMZ/A 2012
MISSION_LEADER	VARCHAR2(50)	Y	Mission leader	Chief scientist / principal investigator; LASTNAME,FIRSTNAME
MISSION_SDATE	DATE	Y	Mission start date (DD/MM/YYYY; UTC)	
MISSION_EDATE	DATE	Y	Mission end date (DD/MM/YYYY; UTC)	
MISSION_INSTITUTE	VARCHAR2(50)	Y	Sponsoring institute	The institute responsible at the time of data collection. Current DFO examples: NAFC, BIO, SABS, MONC, IMI., IOS. "Unknown" is acceptable for historical data
MISSION_PLATFORM	VARCHAR2(50)	Y	Specific platform or vessel	May be vessel name, fishing boat, wharf, various small vessels, multiple ships. Check that name is spelled correctly. "Unknown" is acceptable for historical data
MISSION_PROTOCOL	VARCHAR2(50)	Y	Identify a standard sampling protocol if applicable, e.g., JGOFS, AZMP	A citation should be given if standard protocols were used during the mission. The use of non-standard protocols should be noted and further details provided in the BCCOLLECTOR_COMMENT field
MISSION_GEOGRAPHIC_REGION	VARCHAR2(100)	Y	General geographic region	Examples: Scotian Shelf, lower St. Lawrence Estuary
MISSION_COLLECTOR_COMMENT1	VARCHAR2(200)	Y	Comments on affiliations, collaborations, or other information from collector	Comments from the collector that are pertinent to the entire mission. Generally referring to data collection, analysis, publications, joint missions (more than one institute involved)
MISSION_COLLECTOR_COMMENT2	VARCHAR2(200)	Y	Continuation of the mission_collectors_comment1 for long comments	
MISSION_DATA_MANAGER_COMMENT	VARCHAR2(200)	Y	Comments from the data centre/data manager	Comments from the data manager that are pertinent to the entire mission. Generally referring to data management history (processing steps, edits, special warnings)
EVENT_SDATE	DATE	N	Start date for event (DD/MM/YYYY; UTC)	

Column name	Column data type	Null	Column comment	Best practices: examples (if applicable)
EVENT_EDATE	DATE	Y	End date for event (DD/MM/YYYY; UTC)	Leave blank if unknown
EVENT_STIME	NUMBER(4)	Y	Start time for event (HHMM; UTC)	Leave blank if unknown (e.g., historical data)
EVENT_ETIME	NUMBER(4)	Y	End time for event (HHMM; UTC)	Leave blank if unknown; do not use stime as default
EVENT_MIN_LAT	NUMBER(8,5)	N	Min latitude for event (decimal degrees; +ve north)	Need to calculate from start-end positions of all sampling done at this event (if only one set of coordinates, min = start)
EVENT_MAX_LAT	NUMBER(8,5)	Y	Max latitude for event (decimal degrees; +ve north)	Need to calculate from start-end positions of all sampling done at this event (if the only set of coordinates is the start position then, max = null)
EVENT_MIN_LON	NUMBER(9,5)	N	Min longitude for event (decimal degrees; +ve east)	Need to calculate from start-end positions of all sampling done at this event (if only one set of coordinates, min = start)
EVENT_MAX_LON	NUMBER(9,5)	Y	Max longitude for event (decimal degrees; +ve east)	Need to calculate from start-end positions of all sampling done at this event (if the only set of coordinates is the start position then, max = null)
EVENT_UTC_OFFSET	NUMBER(4,1)	Y	Time zone offset of work region (decimal hours)	All time/dates are in UTC; this field gives the difference between UTC and local time, e.g., EST=+5, AST=+4, NST=+3.5. If unknown, leave blank (NOT ZERO)
EVENT_COLLECTOR_COMMENT1	VARCHAR2(200)	Y	Collector comments on event	Any pertinent comments about event
EVENT_COLLECTOR_COMMENT2	VARCHAR2(200)	Y	Collector comments on event; continued from the event_collectors_comment1 column	
EVENT_DATA_MANAGER_COMMENT	VARCHAR2(200)	Y	Comments from the data centre/data manager	Generally comments related to data management history, e.g., QC
DIS_HEADR_GEAR_SEQ	NUMBER(8)	N	Foreign key to a gear_seq value from the BCGEARS table	See code table
DIS_HEADR_SDATE	DATE	N	Specific start date of collection (DD/MM/YYYY; UTC)	
DIS_HEADR_EDATE	DATE	Y	Specific end date of collection (DD/MM/YYYY; UTC)	Leave blank if unknown; do not use sdate as default
DIS_HEADR_STIME	NUMBER(4)	Y	Specific start time of collection (HHMM; UTC)	Leave blank if unknown (e.g., historical data)
DIS_HEADR_ETIME	NUMBER(4)	Y	Specific end time of collection (HHMM; UTC)	Leave blank if unknown; do not use stime as default
DIS_HEADR_TIME_QC_CODE	VARCHAR2(2)	N	Foreign key to the BCQUALCODES table qualifying the time record	Should be between 0 and 9. Generally 1 = correct. Set flag to 5 if time was corrected; 3 if doubtful, 4 if incorrect. QC 3 and 4 should be avoided! But may be necessary for historical data
DIS_HEADR_SLAT	NUMBER(8,5)	N	Specific start latitude of collection (decimal degrees; +ve north)	
DIS_HEADR_ELAT	NUMBER(8,5)	Y	Specific end latitude of collection (decimal degrees; +ve north)	Leave blank if unknown; do not use slat as default
DIS_HEADR_SLON	NUMBER(9,5)	N	Specific start longitude of collection (decimal degrees; +ve east)	
DIS_HEADR_ELON	NUMBER(9,5)	Y	Specific end longitude of collection (decimal degrees; +ve east)	Leave blank if unknown; do not use slong as default
DIS_HEADR_POSITION_QC_CODE	VARCHAR2(2)	N	Foreign key to the BCQUALCODES table qualifying the position record	Should be between 0 and 9. Generally 1 = correct. Set flag to 5 if time was corrected; 3 if doubtful, 4 if incorrect. QC 3 and 4 should be avoided! But may be necessary for historical data
DIS_HEADR_START_DEPTH	NUMBER(7,2)	N	Minimum collection depth in metres	
DIS_HEADR_END_DEPTH	NUMBER(7,2)	N	Maximum collection depth in metres	Min and max sample depths are generally the same for discrete samples, but may be different if sample is e.g., integrated water column
DIS_HEADR_SOUNDING	NUMBER(5)	Y	Water depth in metres	

Column name	Column data type	Null	Column comment	Best practices: examples (if applicable)
DIS HEADR COLLECTOR_DEPLMT_ID	VARCHAR2(50)	Y	Original deployment ID as provided by the data provider/collector	Often institute mission number + station sequence number. May be the same as EVENT_COLLECTOR_EVENT_ID
DIS HEADR COLLECTOR_SAMPLE_ID	VARCHAR2(50)	N	Individual sample ID, i.e., common identifier label	Unique sample number; sometimes mission//stn//depth
DIS HEADR COLLECTOR	VARCHAR2(50)	Y	Name of the individual who collected the data	Person responsible for the data; the prefix (if present) may indicate the program under which data was collected
DIS HEADR COLLECTOR_COMMENT1	VARCHAR2(200)	Y	Collector comments	May contain information on sample gear, e.g., Rosette: SBE911 + Niskin bottles
DIS HEADR DATA_MANAGER_COMMENT	VARCHAR2(200)	Y	Comments from the data centre/data manager	
DIS HEADR RESPONSIBLE_GROUP	VARCHAR2(50)	Y	The department/division/group responsible for the data record (as opposed to an individual's name)	Collector's institute/organization/university/department
DIS HEADR SHARED_DATA	VARCHAR2(50)	Y	Organization that was sent a copy of the data (i.e., Data may be sent to OSD)	
CREATED_BY	VARCHAR2(30)	N	User ID of the person who uploaded the data	Usually data manager's name
CREATED_DATE	DATE	N	Date that data were uploaded to the edit table	
DATA_CENTER_CODE	NUMBER(2)	N	Code representing the data centre (that uploaded data). FK reference to the BCDATACENTERS table	See code table

(b) BCDISCRETEDATAEDITS (BCD): discrete data

Column name	Column data type	Null	Column comment	Best practices: examples (if applicable) NOTE: the first few fields must match those in the BCS table
DIS DATA NUM	NUMBER(12)	N	Generated sequence number providing unique numeric values for the BCDISCRETEDATAEDITS primary key column	
MISSION_DESCRIPTOR	VARCHAR2(50)	N	Cruise number (or similar)	must match entry in BCS table
EVENT_COLLECTOR_EVENT_ID	VARCHAR2(50)	N	Number assigned to station during data collection	must match entry in BCS table
EVENT_COLLECTOR_STN_NAME	VARCHAR2(50)	Y	Descriptive name of station (e.g., stn27)	must match entry in BCS table
DIS_HEADER_START_DEPTH	NUMBER(7,2)	N	Minimum collection depth in metres	must match entry in BCS table
DIS_HEADER_END_DEPTH	NUMBER(7,2)	N	Maximum sample depth in metres	must match entry in BCS table
DIS_HEADER_SLAT	NUMBER(8,5)	N	Specific latitude of collection (decimal degrees; +ve north)	must match entry in BCS table
DIS_HEADER_SLON	NUMBER(9,5)	N	Specific longitude of collection (decimal degrees; +ve east)	must match entry in BCS table
DIS_HEADER_SDATE	DATE	N	Specific date of collection (DD/MM/YYYY; UTC)	must match entry in BCS table
DIS_HEADER_STIME	NUMBER(4)	Y	Specific time of collection (HHMM; UTC)	must match entry in BCS table
DIS_DETAIL_DATA_TYPE_SEQ	NUMBER(8)	N	Foreign key to the BCDATATYPES table to identify the variable	See code table
DATA_TYPE_METHOD	VARCHAR2(20)	Y	Short name identifying the variable (typically will include the analysis method)	See code table
DIS_DETAIL_DATA_VALUE	NUMBER(10,5)	N	Data value	
DIS_DETAIL_DATA_QC_CODE	VARCHAR2(2)	N	Foreign key to the BCQUALCODES table qualifying the data value	See code table
DIS_DETAIL_DETECTION_LIMIT	NUMBER(11,5)	Y	Detection limit of the observed variable	Include detection limit if available
DIS_DETAIL_DETAIL_COLLECTOR	VARCHAR2(50)	Y	Name of the individual who collected the data	Contact person if there are questions about data collection
DIS_DETAIL_COLLECTOR_SAMP_ID	VARCHAR2(50)	N	Collector's sample ID	must match entry in BCS table
CREATED_BY	VARCHAR2(30)	N	User ID of the person who uploaded the data	must match entry in BCS table
CREATED_DATE	DATE	N	Date that data were uploaded to the edit table	must match entry in BCS table
DATA_CENTER_CODE	NUMBER(2)	N	Code representing the data centre (that uploaded data). FK reference to the BCDATACENTERS table	See code table
DIS_SAMPLE_KEY_VALUE	VARCHAR2(50)	N	FK to the BCDISCRETEDATAEDITS table	

(c) BCPLANKTONSTATNEDITS (BCS): station information for plankton data

Column name	Column data type	Null	Column comment	Best practices: examples (if applicable)
PLANK_SAMPLE_KEY_VALUE	VARCHAR2(50)	N	Generated sequence number providing unique numeric values for the BCPLANKTONSTATNEDITS primary key column	
MISSION_NAME	VARCHAR2(50)	Y	Mission name	Originator's mission number and/or common name(s) for the mission; e.g., IMI-2012-44 Ice/PM/A 2012
MISSION_DESCRIPTOR	VARCHAR2(50)	N	Cruise number (or similar)	Code assigned by OSD, ensures national coordination; e.g., 18HU12044 (see Annex II)
MISSION_LEADER	VARCHAR2(50)	Y	Mission leader	Chief scientist / principal investigator; LASTNAME, FIRSTNAME
MISSION_SDATE	DATE	Y	Mission start date (DD/MM/YYYY; UTC)	
MISSION_EDATE	DATE	Y	Mission end date (DD/MM/YYYY; UTC)	
MISSION_INSTITUTE	VARCHAR2(50)	Y	Sponsoring institute	The institute responsible at the time of data collection. Current DFO examples: NAFC, BIO, SABS, MONC, IMI, IOS. "Unknown" is acceptable for historical data
MISSION_PLATFORM	VARCHAR2(50)	Y	Specific platform or vessel	May be vessel name, fishing boat, wharf, various small vessels, multiple ships. Check that name is spelled correctly. "Unknown" is acceptable for historical data
MISSION_PROTOCOL	VARCHAR2(50)	Y	Identify a standard sampling protocol if applicable, e.g., JGOFS, AZMP	A citation should be given if standard protocols were used during the mission. The use of non-standard protocols should be noted and further details provided in the BCCOLLECTOR_COMMENT field
MISSION_GEOGRAPHIC_REGION	VARCHAR2(100)	Y	General geographic region	Examples: Scotian Shelf, lower St. Lawrence Estuary
MISSION_COLLECTOR_COMMENT	VARCHAR2(200)	Y	Comments on affiliations, collaborations, or other information from collector	Comments from the collector that are pertinent to the entire mission. Generally referring to data collection, analysis, publications, joint missions (more than one institute involved)
MISSION_MORE_COMMENT	VARCHAR2(200)	Y	Continuation of the mission collectors comment column for long comments	
MISSION_DATA_MANAGER_COMMENT	VARCHAR2(200)	Y	Comments from the data centre/data manager	Comments from the data manager that are pertinent to the entire mission. Generally referring to data management history (processing steps, edits, special warnings)
EVENT_SDATE	DATE	N	Start date for event (DD/MM/YYYY; UTC)	
EVENT_EDATE	DATE	Y	End date for event (DD/MM/YYYY; UTC)	Leave blank if unknown
EVENT_STIME	NUMBER(4)	Y	Start time for event (HHMM; UTC)	Leave blank if unknown (e.g., historical data)
EVENT_ETIME	NUMBER(4)	Y	End time for event (HHMM; UTC)	Leave blank if unknown; do not use stime as default
EVENT_MIN_LAT	NUMBER(8,5)	N	Min latitude for event (decimal degrees; +ve north)	Need to calculate from start-end positions of all sampling done at this event (if only one set of coordinates, min = start)
EVENT_MAX_LAT	NUMBER(8,5)	Y	Max latitude for event (decimal degrees; +ve north)	Need to calculate from start-end positions of all sampling done at this event (if the only set of coordinates is the start position then, max = null)
EVENT_MIN_LON	NUMBER(9,5)	N	Min longitude for event (decimal degrees; +ve east)	Need to calculate from start-end positions of all sampling done at this event (if only one set of coordinates, min = start)
EVENT_MAX_LON	NUMBER(9,5)	Y	Max longitude for event (decimal degrees; +ve east)	Need to calculate from start-end positions of all sampling done at this event (if the only set of coordinates is the start position then, max = null)
EVENT_COLLECTOR_STN_NAME	VARCHAR2(50)	Y	Descriptive name of station	Can be descriptive station name, or consec//, e.g., TASO1

Column name	Column data type	Null	Column comment	Best practices: examples (if applicable)
EVENT_COLLECTOR_EVENT_ID	VARCHAR2(50)	N	Number assigned to station during data collection (may not be applicable)	Unique identifier for the station, may be mission//+stn name+consecutive//; e.g., IMI.2012044TASOI_66 (mission IMI_2012-44, station TASOI, consecutive 66)
EVENT_UTC_OFFSET	NUMBER(4,1)	Y	Time zone offset of work region (decimal hours)	All time/dates are in UTC; this field gives the difference between UTC and local time, e.g., EST=-5, AST=-4, NST=-3.5. If unknown, leave blank (NOT ZERO)
EVENT_COLLECTOR_COMMENT	VARCHAR2(200)	Y	Collector comments on event	Any pertinent comments about event
EVENT_MORE_COMMENT	VARCHAR2(200)	Y	Collector comments on event; continued from the event_collectors_comment1 column	
EVENT_DATA_MANAGER_COMMENT	VARCHAR2(200)	Y	Comments from the data centre/data manager	Generally comments related to data management history, e.g., QC
PL_HEADR_GEAR_SEQ	NUMBER(8)	N	Reference to a gear_seq value from the BCGEARS table	See code table
PL_HEADR_SDATE	DATE	N	Specific start date of collection (DD/MM/YYYY; UTC)	
PL_HEADR_EDATE	DATE	Y	Specific end date of collection (DD/MM/YYYY; UTC)	Leave blank if unknown; do not use sdate as default
PL_HEADR_STIME	NUMBER(4)	Y	Specific start time of collection (HHMM; UTC)	Leave blank if unknown (e.g., historical data)
PL_HEADR_ETIME	NUMBER(4)	Y	Specific end time of collection (HHMM; UTC)	Leave blank if unknown; do not use stime as default
PL_HEADR_PHASE_OF_DAYLIGHT	VARCHAR2(15)	Y	Used to identify the time of day (i.e., day, night, twilight)	Leave blank if unknown
PL_HEADR_SLAT	NUMBER(8,5)	N	Specific start latitude of collection (decimal degrees; +ve north)	
PL_HEADR_ELAT	NUMBER(8,5)	Y	Specific end latitude of collection (decimal degrees; +ve north)	Leave blank if unknown; do not use slat as default
PL_HEADR_SLON	NUMBER(9,5)	N	Specific start longitude of collection (decimal degrees; +ve east)	
PL_HEADR_ELO	NUMBER(9,5)	Y	Specific end longitude of collection (decimal degrees; +ve east)	Leave blank if unknown; do not use slong as default
PL_HEADR_TIME_QC_CODE	VARCHAR2(2)	N	Foreign key to the BCQUALCODES table qualifying the time record	Should be between 0 and 9. Generally 1 = correct. Set flag to 5 if time was corrected; 3 if doubtful, 4 if incorrect. QC_3 and 4 should be avoided! But may be necessary for historical data
PL_HEADR_POSITION_QC_CODE	VARCHAR2(2)	N	Foreign key to the BCQUALCODES table qualifying the position record	Should be between 0 and 9. Generally 1 = correct. Set flag to 5 if position was corrected; 3 if doubtful, 4 if incorrect. QC_3 and 4 should be avoided! But may be necessary for historical data
PL_HEADR_START_DEPTH	NUMBER(7,2)	N	Minimum collection depth in metres	Min and max depths may be the same (e.g., sample from water bottle, stationary submersible pump)
PL_HEADR_END_DEPTH	NUMBER(7,2)	N	Maximum collection depth in metres	Min and max depths may be the same (e.g., sample from water bottle, stationary submersible pump)
PL_HEADR_SOUNDING	NUMBER(5)	Y	Water depth in metres	
PL_HEADR_VOLUME	NUMBER(7,3)	Y	Volume of water used to calculate abundance per unit volume (cubic metres); e.g., vol. of water filtered through net or sample volume (phytoplankton or other microorganisms)	

Column name	Column data type	Null	Column comment	Best practices: examples (if applicable)
PL_HEADR_VOLUME_METHOD_SEQ	NUMBER(8)	N	Volume method code from the BCVOLUMEMETHOD table	See code table
PL_HEADR_LRG_PLANKTON_REMOVED	CHAR(1)	Y	Flag indicating that the procedure included removal of large plankton (Y / N)	Yes/No entry to indicate if procedure included removal of large plankton e.g., jellyfish, from the sample prior to preservation
PL_HEADR_MESH_SIZE	NUMBER(6)	Y	Mesh size in microns	
PL_HEADR_COLLECTION_METHOD_SEQ	NUMBER(8)	N	Collection method code from the BCCOLLECTIONMETHODS table	See code table
PL_HEADR_COLLECTOR_DEPLMT_ID	VARCHAR2(50)	Y	Original label as provided by the data provider/collector	UniqueID
PL_HEADR_COLLECTOR_SAMPLE_ID	VARCHAR2(50)	Y	Individual sample ID, e.g., common identifier label	Unique sample number; sometimes mission//+stn#
PL_HEADR_PROCEDURE_SEQ	NUMBER(8)	N	Procedure code from the BCPROCEDURES table	See code table
PL_HEADR_PRESERVATION_SEQ	NUMBER(8)	N	Preservation code from the BCPRESERVATIONS table	See code table
PL_HEADR_STORAGE_SEQ	NUMBER(8)	N	Storage code from the BCSTORAGES table	See code table
PL_HEADR_COLLECTOR	VARCHAR2(50)	Y	Name of the individual who collected the data	Contact person if there are questions about data collection
PL_HEADR_COLLECTOR_COMMENT	VARCHAR2(200)	Y	Collector comments	
PL_HEADR_METERS_SQD_FLAG	CHAR(1)	Y	Meters squared flag [Y or N] indicates whether the value can be used for standing stock calculation	
PL_HEADR_DATA_MANAGER_COMMENT	VARCHAR2(200)	Y	Comments from the data centre/data manager	
PL_HEADR_RESPONSIBLE_GROUP	VARCHAR2(50)	Y	The department/division/group responsible for the data record (as opposed to an individual's name)	Collector's institute/organization/university/department
PL_HEADR_SHARED_DATA	VARCHAR2(50)	Y	Organization that may receive a copy of the data (i.e., Data may be sent to OSD)	
CREATED_BY	VARCHAR2(30)	N	User ID of the person who uploaded the data	Usually data manager's name
CREATED_DATE	DATE	N	Date that data were uploaded to the edit table	
DATA_CENTER_CODE	NUMBER(2)	N	Code representing the data centre (that uploaded data / from which data may be obtained). FK reference to the BCDATACENTERS table	See code table
BATCH_SEQ	NUMBER(10)	Y	Stores the user specified batch number or job number in order to group/categorize a dataset that is in the process of being validated. This will enable a user to process more than one dataset at one time	For example: AAAAMMMSSS. AAAA for the year the mission was done; MMM mission number; SSS sequence. For zooplankton data of the IML-12-44 mission, it could be: 2012044001 and for the phytoplankton dataset it could be 2012044002

(d) BCPLANKTONDATAEDITS (BCD): plankton data

Column name	Column data type	Null	Column comment	Best practices: examples (if applicable)
PLANK_DATA_NUM	NUMBER(14)	N	Generated sequence number providing a unique numeric values for the BCPLANKTONDATAEDITS primary key column	
PLANK_SAMPLE_KEY_VALUE ¹	VARCHAR2(50)	N	Unique values from the BCPLANKTONSTATNEDITS primary key column	must match entry in BCS table
MISSION_DESCRIPTOR	VARCHAR2(50)	N	Cruise number (or simliar)	must match entry in BCS table
EVENT_COLLECTOR_EVENT_ID	VARCHAR2(50)	N	Number assigned to station during data collection (may not be applicable)	must match entry in BCS table
EVENT_COLLECTOR_STN_NAME	VARCHAR2(50)	Y	Descriptive name of station (e.g., stn27)(may not be applicable)	must match entry in BCS table
PL_GEN_NATIONAL_TAXONOMIC_SEQ ¹	NUMBER(14)	N	Species code from the BCNATNL TAXONCODES table	See code table
PL_GEN_COLLECTOR_TAXONOMIC_ID	VARCHAR2(20)	Y	Originator's or collector's code	
PL_GEN_LIFE_HISTORY_SEQ ¹	NUMBER(8)	N	Life stage code from the BCLIFEHISTORIES table	See code table
PL_GEN_TROPHIC_SEQ	NUMBER(8)	N	Trophic stage code from the BCTROPHICDESCRIPTORS table	See code table
PL_GEN_MIN_SIEVE ¹	NUMBER(8,4)	Y	Retention filter size (mm)	If sample size fractionated, minimum sieve size (this had been used in the past to indicate minimum size, e.g., copepoda egg (0.135-0.165 mm))
PL_GEN_MAX_SIEVE ¹	NUMBER(8,4)	Y	Largest sieve used (mm)	If sample size fractionated, maximum sieve size (this had been used in the past to indicate the maximum size, e.g., copepoda egg (0.135-0.165 mm))
PL_GEN_MODIFIER ¹	VARCHAR2(50)	Y	Information that complements the description or qualifies the organism in addition to its name	Organism size, colour, other details from ID sheet. This is also the field to populate with name qualifiers such as sp, spp, ?, unidentified, sp. A, fragments, damaged, aff, cf.
PL_GEN_SPLIT_FRACTION	NUMBER(5,4)	Y	Fraction of sample (0.0 - 1.0)	Proportion of the sample analyzed
PL_GEN_SEX_SEQ ¹	NUMBER(8)	N	Sex code from the BCSEXES lookup table	See code table
PL_GEN_COUNTS	NUMBER(15,3)	Y	Number of organisms counted	
PL_GEN_COUNT_PCT	NUMBER(6,3)	Y	Percentage of the specified plankton	
PL_GEN_WET_WEIGHT	NUMBER(9,4)	Y	Wet weight of organisms (grams)	
PL_GEN_DRY_WEIGHT	NUMBER(9,4)	Y	Dry weight of organisms (grams)	
PL_GEN_BIO_VOLUME	NUMBER(8,3)	Y	Settled volume of organisms (mL)	
PL_GEN_PRESENCE	CHAR(1)	Y	Indicates presence or absence of organism(s) if not counted (Y / N)	
PL_GEN_COLLECTOR_COMMENT	VARCHAR2(200)	Y	Collector comments	
PL_GEN_DATA_MANAGER_COMMENT	VARCHAR2(200)	Y	Comments from the data centre/data manager	
PL_GEN_SOURCE	VARCHAR2(30)	N	Tracks the name of the individual or source of the sampled data (subsampled data) if a particular sample is to be reanalyzed	
PL_FREQ_DATA_TYPE_SEQ ²	NUMBER(8)	N	Auto-generated sequence number to reference a measurement parameter. Foreign key to the BCDATATYPES table	

Column name	Column data type	Null	Column comment	Best practices: examples (if applicable)
PL_FREQ_UPPER_BIN_SIZE ²	NUMBER(6,3)	N	Upper limit of bin size	
PL_FREQ_LOWER_BIN_SIZE ²	NUMBER(6,3)	N	Lower limit of bin size	
PL_FREQ_BUG_COUNT ²	NUMBER(6)	N	Number of organisms in extracted subsample	
PL_FREQ_BUG_SEQ ²	NUMBER(6)	N	Auto-generated sequential number to track individual organisms	
PL_FREQ_DATA_VALUE ²	NUMBER(10,5)	N	Data value	
PL_FREQ_DATA_QC_CODE ²	VARCHAR2(2)	N	Quality control code for data. Foreign key to the BCQUALCODES table	Should be between 0 and 9
PL_FREQ_DETAIL_COLLECTOR ²	VARCHAR2(50)	Y	Name of the collector (individual) for the data	
PL_DETAIL_DATA_TYPE_SEQ ²	NUMBER(8)	N	Auto-generated sequence number to reference a measurement parameter. Foreign key to the BCDATATYPES table	
PL_DETAIL_DATA_VALUE ²	NUMBER(10,5)	N	Data value	
PL_DETAIL_DATA_QC_CODE ²	VARCHAR2(2)	N	Quality control code for data. Foreign key to the BCQUALCODES table	Should be between 0 and 9
PL_DETAIL_DETAIL_COLLECTOR ²	VARCHAR2(50)	Y	Name of the collector (individual) for the data	
PL_INDIV_DATA_TYPE_SEQ ²	NUMBER(8)	N	Auto-generated sequence number to reference a measurement parameter. Foreign key to the BCDATATYPES table	
PL_INDIV_BUG_SEQ ²	NUMBER(6)	N	Auto-generated sequential number to track individual organisms	
PL_INDIV_DATA_VALUE ²	NUMBER(10,5)	N	Data value	
PL_INDIV_DATA_QC_CODE ²	VARCHAR2(2)	N	Quality control code for data. Foreign key to the BCQUALCODES table	Should be between 0 and 9
PL_INDIV_DATA_COLLECTOR ²	VARCHAR2(50)	Y	Name of the collector (individual) for the data	
CREATED_BY	VARCHAR2(30)	N	User ID of the person who uploaded the data	must match entry in BCS table
CREATED_DATE	DATE	N	Date that data were uploaded to the odit table	must match entry in BCS table
DATA_CENTER_CODE	NUMBER(2)	N	Code representing the data centre (that uploaded data). FK reference to the BCDATACENTERS table	See code table
BATCH_SEQ	NUMBER(10)	Y	Stores the user specified batch number or job number in order to group/categorize a dataset that is in the process of being validated. This will enable a user to process more than one dataset at one time	For example: AAAAMMMSSS. AAAA for the year the mission was done; MMM mission number; SSS sequence. For zooplankton data of the IM1-12-44 mission, it could be: 2012044001 and for the phytoplankton dataset it could be 2012044002

1: These seven fields together must make a unique combination; if not, the database considers entries as duplicates and will only load one occurrence.

2: If there are no data for FREQUENCY, DETAIL, of INDIVIDUAL, these fields may be left blank. However, if data exist, null / not null rules must be respected.